Universidad de Salamanca

Departamento de Estadística

Máster en Análisis Avanzado de Datos Multivariantes y Big Data

Trabajo Fin de Máster

MANOVA Bootstrap basado en distancias

Laura Vicente González– 70916732-G

Trabajo supervisado por

Dr. José Luis VICENTE-VILLARDÓN

16 de julio de 2019

Departamento de Estadística

Universidad de Salamanca

José Luis VICENTE-VILLARDÓN

Profesor Titular del Departamento de Estadística de la Universidad de Salamanca

CERTIFICA que **Dª Laura Vicente González** ha realizado en la Universidad de Salamanca, bajo su dirección, el trabajo para optar al título del Máster en Análisis Avanzado de Datos Multivariantes y Big Data que presenta con el título de **MANOVA Bootstrap basado en distancias**, autorizando expresamente su lectura y defensa.

Y para que conste, firma el presente certificado en Salamanca a 16 de julio de 2019.

José Luis VICENTE-VILLARDÓN

Departamento de Estadística

Universidad de Salamanca

Laura Vicente González

Alumna del Máster en Análisis Avanzado de Datos Multivariantes y Big Data

PRESENTA en la Universidad de Salamanca, el trabajo para optar al título del Máster en Análisis Avanzado de Datos Multivariantes y Big Data con el título de **MANOVA Bootstrap basado en distancias** realizado bajo la dirección de José Luis VICENTE-VILLARDÓN.

Y para que conste, firma el presente documento en Salamanca a 16 de julio de 2019.

Laura Vicente González

Publicado en 16 de julio de 2019 por

Laura Vicente González

laura20vg@usal.es

Universidad de Salamanca

Agradecimientos

Quiero agradecir a toda mi familia y al Departamento de Estadística de la Universidad de Salamanca por las enseñanzas y el amparo incondicional durante todos estos años, sobre todo en este último año en el que he realizado este Máster en Análisis Avanzado de Datos Multivariantes y Big Data. También agradecer a mis compañeros por el apoyo y la ayuda recibida.

Resumen

Es cada vez más frecuente encontrar grandes matrices de datos con un número elevado de variables, incluso mayor que el número de individuos. Cuando se trata de establecer diferencias significativas entre grupos deberían utilizarse los métodos de contraste multivariantes para controlar el riesgo tipo I. El método más popular es el Análisis Multivariante de la Varianza (MANOVA) que puede considerarse como un caso particular del Modelo Lineal General Multivariante (MLGM). Normalmente el MANOVA se acompaña de una representación gráfica (Análisis Canónico) para ayudar con la interpretación en caso de que se rechace la hipótesis nula de igualdad de vectores de medias. El problema del MANOVA es que tiene condiciones de aplicación muy restrictivas, los datos tienen que tener distribuciones normales multivariantes y la estructura de variación y covariación tiene que ser la misma en todos los grupos; además, el número de variables tiene que ser mucho menor que el número de individuos para que el modelo sea adecuado. En muchos casos prácticos estas condiciones no se cumplen y es necesario recurrir a métodos no paramétricos. Utilizaremos como alternativa el PERMANOVA y el BOOTMANOVA. En este trabajo describiremos el PERMANOVA (en el segundo capítulo) haciendo referencia a su relación con el MANOVA y el MLGM que describimos en el primer capítulo. También se desarrollará en el capítulo 2, como alternativa al PERMANOVA, el BOOTMANOVA, tiene su fundamento en el MANOVA basado en distancias y emplea técnicas bootstrap para hacer la estimación de la distribución muestral. En el tercer capítulo se explicarán las técnicas de representación para datos continuos y binarios asociados al PERMANOVA y al BOOTMANOVA, concretamente el Análisis de Coordenadas sobre los centroides y un Análisis Canónico Bootstrap. Finalmente aplicaremos las técnicas mencionadas a cuatro conjuntos de datos genéticos.

Keywords: MANOVA, Bootstrap, PERMANOVA, BOOTMANOVA, Análisis Canónico

Índice general

Capítulo 1

Introducción

La Estadística se define como la ciencia que utiliza conjuntos de datos numéricos para obtener, a partir de ellos, inferencias sobre las poblaciones en las que fueron recogidos, basadas en ocasiones en el cálculo de las probabilidades. Por tanto, es necesario organizar y almacenar datos que puede ser sometidos a estudio. A lo largo de la historia, el almacenamiento de datos ha ido evolucionando a gran velocidad. Los primeros sistemas de almacenamiento de datos fueron las tarjetas perforadas en la década de 1960 que podían contener en torno a 90 caracteres, en la actualidad, cuando han trascurrido menos de 60 años desde que se fabricaron los primeros, este tipo de dispositivos físicos de almacenamiento están perdiendo importancia y se almacena un gran número de GB de datos en la nube para tener acceso desde cualquier dispositivo.

Al evolucionar a tanta velocidad, la estadística se ha visto obligada a desarrollar nuevas técnicas tanto para la recogida de datos como para el análisis de los mismos.

El desarrollo de las nuevas tecnologías ha provocado que se generen datos continuamente, a través de multitud de aplicaciones o con la extracción de datos que antes no era posible obtener. Esto nos sitúa en la que denominan como la era del big data. La necesidad de analizar estos grandes conjuntos de datos ha puesto en auge las técnicas estadísticas multivariantes que permiten trabajar con multitud de variables de forma simultánea. En algunas ocasiones, dichos conjuntos de datos el número de individuos que se muestrean es inferior al número de variables medidas. Esto ocurre entre otros en los datos genómicos, en los que se recogen, por ejemplo a través de microarrays, secuencias de ADN o ARN en los que se mide la expresión génica de miles de genes para unos pocos individuos. Generalmente, estos individuos son clasificados en grupos; algunos

de ellos pueden ser por el sexo, por si padece una enfermedad o no, por el lugar donde se ha recogido la muestra, por el tratamiento que han recibido o por cualquier otra variable categórica que pueda servir como variable de agrupación.

Uno de los objetivos más comunes, utilizando las variables de agrupación mencionadas anteriormente, es buscar diferencias entre los grupos de uno de los factores o teniendo en cuenta varios de ellos. La técnica de Análisis Multivariante más conocida, y por lo tanto la más utilizada, para realizar dicha comparación es el Análisis Multivariante de la Varianza (MANOVA). Esta técnica está basada en un Modelo Lineal General Multivariante (MGLM) y consiste en la partición de la variabilidad de los datos buscando las diferencias significativas existentes entre los grupos de individuos. Para poder aplicar el MANOVA, los datos estudiados deben cumplir tres condiciones fundamentales:

- Seguir una distribución normal multivariante.
- Presentar homocedasticidad, es decir, que las matrices de varianzas y covarianzas de los grupos sean iguales.
- El número de individuos sea menor que el número de variables ya que en caso contrario, al estar basado en un MLGM, algunos de los cálculos necesarios para la estimación de los parámetros no es posible realizaros y los estimadores no están definidos.

A pesar de que el MANOVA está ampliamente extendido, la mayor parte de los investigadores se limitan a utilizar las técnicas univariantes separadas para cada una de las variables, generalmente un Análisis de la Varianza (ANOVA) o sus equivalentes no paramétricos. El caso de los datos genómicos puede servir de ejemplo, un gran número de investigadores aplica las técnicas paramétricas univariantes, a pesar de que la distribución de los datos sea marcadamente asimétrica. Las técnicas multivariantes para la búsqueda de la significación de las diferencias entre grupos, seguramente por tener una mayor complejidad que las univariantes y por los supuestos básicos que deben cumplir, no han recibido el reconocimiento que se merecen, aunque en muchos casos fuera más correcta su aplicación que la de la técnica univariante análoga.

Son muchos los conjuntos de datos multivariantes que no cumplen las hipótesis básicas para poder realizar un MANOVA a la población de estudio (Xu and Cui, 2008), por ello han sido desarrolladas técnicas multivariantes alternativas que permitan su estudio en estos casos. Las

técnicas desarrolladas hasta el momento no han recibido una gran atención en la bibliografía especializada, la mayor parte de ellas se pueden encontrar en el ámbito de la Ecología, que fue precursora de alguna de ellas. En este trabajo se presentará una nueva alternativa no desarrollada hasta la actualidad para datos con este tipo de características.

Uno de los casos menos estudiados es aquel en el que el número de individuos de la muestra seleccionada (I) es menor que el número de variables respuesta (J), sin embargo en la literatura se pueden encontrar diferentes técnicas que permiten la realización de dichos estudios. Algunas de estas técnicas son ANOSIM (Clarke, 1993), la prueba de Mantel o el Análisis Permutacional Multivariante de la Varianza (PERMANOVA) (Anderson, 2001; McArdle and Anderson, 2001) que consiste en la realización de un Análisis de Permutaciones combinado con un Análisis Multivariante de la Varianza para obtener los resultados buscados. Esta última técnica será descrita brevemente en la sección 3.3 de nuestro trabajo, también se podrá encontrar una serie de ejemplos prácticos de la misma. Las técnicas citadas anteriormente tienen aplicaciones en diversos campos, como puede ser el estudio de los patrones ecológicos en ensambles (Chapman and Underwood, 1999), la inflamación intestinal enfocada a la actividad inductora de cáncer (Arthur et al., 2012), la genética del paisaje (Manel et al., 2003), los patrones en las migraciones de las aves (Flather Curtis H. and Sauer John R., 1996) o la prueba de asociación para la composición de la comunidad microbiana (Tang et al., 2016).

El objetivo principal de este trabajo reside en la obtención de una técnica similar a las anteriores, pero empleando técnicas Bootstrap en su procedimiento. El bootstrap es una técnica de creación bastante reciente (Efron, 1979), su base reside en el remuestreo con reposición, a diferencia del análisis de permutaciones (Neyman and S., 1923) que emplea el remuestreo sin reposición. Existen multitud de artículos en los que se emplean ambas técnicas o incluso se comparan. Dependiendo de los supuestos y la finalidad es más conveniente la elección de uno u otro.
En el caso de tener un gran número de datos, el análisis de permutaciones es más complejo ya que, para que el análisis fuera completo sería necesario obtener todas las permutaciones posibles de los datos, generalmente esto no es posible y se coge una muestra lo suficientemente grande como para que pueda ser representativa. Sin embargo, en el caso del bootstrap no existe esta problemática, ya que es el investigador el que, en todo momento, elige el número de remuestreos que desea realizar.

Existen multitud de documentos en la bibliografía que emplean ambas técnicas, en algunos casos son complementarias, en otros pretende hacer una comparación entre ellas (ter Braak, 1992; Præstgaard, 1995; Cheng and Palmer, 2013), sin embargo, no hemos encontrado que trabaje el método bootstrap con la matriz de distancias.

El segundo objetivo es desarrollar una representación gráfica similar al Análisis Canónico (comúnmente utilizado acompañando al MANOVA paramétrico) para ilustrar gráficamente los resultados del BOOTMANOVA. Las regiones de confianza para las medias se calcularán mediante remuestreo bootstrap dentro de cada uno de los grupos.

En el capítulo 2 desarrollaremos el MLGM tradicional en el que se basa el MANOVA descrito a continuación, también nos permitirá determinar la notación a utilizar a lo largo de todo el trabajo. Seguidamente en el capítulo 3 se describirán las dos técnicas centrales del trabajo, es decir los MANOVAs basados en distancias; se realizará un resumen del cálculo de las distancias, una explicación del PERMANOVA descrito por Anderson (2001); McArdle and Anderson (2001) y el desarrollo de la nueva técnica propuesta denominada BOOTMANOVA. El capítulo 4 contiene una representación gráfica en baja dimensión desarrollada tanto para el PERMANOVA como para el BOOTMANOVA, dichos gráficos incluyen la representación de una región de confianza también basada en bootstrap. El último capítulo (capítulo 5) contiene cuatro casos prácticos que permiten ilustrar las técnicas.

Las aplicaciones prácticas contienen los resultados tanto para el PERMANOVA como para el BOOTMANOVA y las representaciones gráficas asociadas a los mismos con la intención de buscar diferencias significativas entre los grupos de individuos. Todas las matrices se corresponden con datos genómicos.

Las bases utilizadas son bases públicas extraídas de la plataforma Gene Expression Omnibus (GEO) en los casos correspondientes a los ejemplos cuantitativos. Los tres primeros casos, los obtenidos de dicha plataforma, contienen expresión génica medida a través de intensidad lumínica a partir de microarrays. El último ejemplo corresponde con un proyecto internacional denominado HapMap (International HapMap Consortium and others, 2005) que contiene datos binarios para la diferenciación de individuos de razas distintas a partir de la presencia o ausencia del SNP.

Los detalles de cada una de las bases de datos se encuentran en el capítulo 5.

Toda la aplicación se ha realizado con el software *R* a través de un paquete de desarrollo propio con la finalidad de realizar los cálculos para ambas técnicas.

Capítulo 2

MANOVA

2.1. Introducción

Dentro de la Estadística, una parte con gran relevancia es el estudio de las relaciones entre dos conjuntos de datos. Estos pueden ser univariantes o multivariantes y tener, o no, papeles simétricos. En este caso trabajaremos con conjuntos de datos multivariantes cuyos papeles no serán simétricos. Uno de los grupos contendrá las variables respuesta que vendrán explicadas por las variables predictoras recogidas en el otro conjunto. Un caso particular a tener en cuenta, será aquel en el que los predictores corresponden con un conjunto de variables cualitativas.

En el caso en el que obtenemos una única variable respuesta, empleamos uno de los métodos clásicos, el Análisis de la Varianza (ANOVA), en caso de disponer de más de una variable respuesta, es decir, se trata de una respuesta multivariante, debemos emplear modelos más complejos, como es el Análisis Multivariante de la Varianza (MANOVA). En ambos casos, se emplean técnicas que englobamos dentro de los Modelos Lineales Generales (univariantes o multivariantes).

En la actualidad, la mayor parte de los experimentos realizados obtienen una repuesta multivariante y un número de mediciones elevado. La práctica habitual de los investigadores en estos casos, consiste en realizar un Análisis Univariante de la Varianza de cada una de las variables por separado, empleando métodos de corrección por comparaciones múltiples para evitar los falsos positivos.

Sin embargo, esta no es la forma óptima de realizarlo, debido a varios puntos. Uno de los más importantes es que la utilización de todas las variables simultáneamente, en gran parte de los casos, aporta información que individualmente no se puede observar, y evita inferencias erró-

neas por tener en cuenta las dependencias entre variables. Otro punto por el que evitar el uso de cada una de las variables por separado, es el control del riesgo Tipo I, se realiza considerando que todas las variables son independientes, pero pueden no serlo. Por último, se puede destacar que en caso de que alguna de las variables no sea significativa, puede existir una combinación de estas que sí que lo sea y pase desapercibida.

La técnica mas indicada para eliminar este tipo de errores es el Análisis Multivariante de la Varianza (MANOVA). Esta técnica permite, a través de combinaciones lineales de las variables observadas, hacer máxima la F de Snedecor univariante. Otra forma de entender el MANOVA es como un Modelo Lineal General Multivariante (MGLM), emplearemos esta última forma para el desarrollo del trabajo.

2.2. Modelo Lineal General Multivariante

Dadas dos matrices $\mathbf{X}_{(IxL)}$ e $\mathbf{Y}_{(IxJ)}$ con L y J variables medidas en I individuos, se pretende estudiar la relación entre las respuestas $\mathbf{Y}$ y las variables explicativas $\mathbf{X}$, que normalmente se consideran constantes y no aleatorias, es decir, determinadas por el investigador.
Así, es posible construir el modelo de la forma

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U} \tag{2.1}$$

donde $\mathbf{B}_{(L\times J)}$ es la matriz que contiene los parámetros de regresión desconocidos y $\mathbf{U}$ el conjunto de errores aleatorios no observables, cada uno de ellos con media 0 y una matriz de covarianzas común $\mathbf{\Sigma}$. Se describe como extensión del Modelo Lineal General Univariante, comúnmente utilizado en multitud de aplicaciones.

Se puede afirmar que, si $\mathbf{X}$ es una matriz de rango completo, los estimadores de $\mathbf{B}$ existen y vienen dados por la expresión

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{2.2}$$

Los estimadores obtenidos en el caso multivariante, se puede observar, que son equivalentes a los que se obtendrían en el modelo univariante para cada una de las variables por separado. El interés del modelo multivariante no reside en las estimaciones, sino en el contraste del modelo

global.

Para este tipo de modelos, es interesante considerar la hipótesis de la forma

$$\mathbf{\Omega} = \mathbf{CBM} = \mathbf{0} \tag{2.3}$$

donde $\mathbf{C}_{S\times L}$ es una matriz de rango S que contiene algún conjunto de combinaciones lineales de las columnas de $\mathbf{X}$, como pueden ser un conjunto de contrastes de las medias en el caso de un diseño experimental. $\mathbf{M}_{J\times R}$ corresponde a la matriz de rango R que permite realzar contrastes a través de las combinaciones lineales de las variables respuesta.
La forma clásica del modelo omite la matriz de contrastes y la matriz $\mathbf{M}$ corresponde a la identidad ($\mathbf{I}$).

El estimador de varianza mínima $\mathbf{\Omega}$ se define como

$$\hat{\mathbf{\Omega}} = \mathbf{C}\hat{\mathbf{B}}\mathbf{M} = \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{M}$$

.

Será necesario definir también $\mathbf{R}$, que tiene relación con la inversa de la matriz de covarianzas entre las variables predictoras, $\mathbf{E}$, la matriz de covarianzas y productos cruzados del error, que está relacionada con $\mathbf{\Sigma}$, y $\mathbf{H}$ que corresponde a la matriz de covarianzas y productos cruzados asociados a la hipótesis.

$$\mathbf{R} = \mathbf{C}\,(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T \tag{2.4}$$

$$\mathbf{E} = \mathbf{M}^T\mathbf{Y}^T[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Y}\mathbf{M} \tag{2.5}$$

$$\mathbf{H} = \mathbf{M}^T\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{M} = \hat{\mathbf{\Omega}}^{\mathbf{T}}\mathbf{R}^{-\mathbf{1}}\hat{\mathbf{\Omega}} \tag{2.6}$$

Para realizar el contraste de la hipótesis (2.3) se empleará el principio de unión– intersección de Roy (1953). Este principio dice que la hipótesis multivariante es verdadera si y solo si las hipótesis univariantes $\mathbf{\Omega a} = \mathbf{CBMa} = \mathbf{0}$ se verifican para todos los vectores no nulos $\mathbf{a}$. Para cada una de las hipótesis univariantes, el estadístico asociado será

$$F(\mathbf{a}) = \frac{(I-L)}{S}\frac{\mathbf{a}^T\mathbf{H}\mathbf{a}}{\mathbf{a}^T\mathbf{E}\mathbf{a}}$$

La hipótesis multivariante (2.3) será aceptada para un nivel de significación α si para todo $\mathbf{a}$ no nulo

$$\bigcap_{\mathbf{a}}[F(\mathbf{a}) \leq F_{\alpha;S,I-L}]$$

Por tanto la región de aceptación puede ser definida como

$$\max_{\mathbf{a}} F(\mathbf{a}) \leq F_{\alpha;S,I-L}$$

Si se introduce en el denominador la restricción ($\mathbf{a}^T\mathbf{E}\mathbf{a} = 1$), se puede demostrar, a través de los Multiplicadores de Lagrange, que el máximo es proporcional a la raíz mayor de la ecuación definida

$$|\mathbf{H} - \lambda\mathbf{E}| = 0 \tag{2.7}$$

con las matrices **H** y **E** citadas anteriormente (2.5 y 2.6).
Las raíces no nulas de la ecuación anterior (2.7) serán las mismas que los valores propios (raíces características) de

$$\mathbf{H}\mathbf{E}^{-1} \tag{2.8}$$

Un gran número de técnicas multivariantes empleadas en la actualidad, utilizan estas raíces características o una función de ellas en sus estadísticos de contraste. Algunas de ellas pueden encontrarse en Morrison (2005); Mardia et al. (2009); Seber (2009).

El primer vector propio de la ecuación (2.8) corresponde con el vector de coeficientes **a** que maximiza la $F - ratio$. Estos coeficientes generan una combinación lineal de las variables respuesta explicadas, lo mejor posible, a partir de las predictoras en una regresión múltiple. Los siguientes vectores propios explicarían, de forma progresiva el máximo de la parte no explicada de la combinación anterior.
Podrían emplearse estas combinaciones lineales para realizar una representación gráfica de los individuos, de forma análoga a un Análisis de Componentes Principales, aunque buscan las combinaciones lineales mejor explicadas por las predictoras, en vez de buscar las combinaciones que explican la mayor parte de la variabilidad.

Usando la matriz (2.8), que generalmente no es simétrica, los vectores pueden obtenerse de la ecuación

$$\left(\mathbf{H}\mathbf{E}^{-1} - \lambda\right)\mathbf{a} = \mathbf{0} \tag{2.9}$$

También será posible realizar la misma descomposición utilizando la matriz simétrica

$$\mathbf{E}^{-1/2}\mathbf{H}\,\mathbf{E}^{-1/2} \tag{2.10}$$

Empleando esta matriz, la ecuación (2.9) puede ser descrita como $[\mathbf{E}^{-1/2}\mathbf{H}\mathbf{E}^{-1/2} - \lambda\mathbf{I}]\mathbf{E}^{1/2}\mathbf{a} = 0$ o $[\mathbf{E}^{-1/2}\mathbf{H}\mathbf{E}^{-1/2} - \lambda\mathbf{I}]\mathbf{v} = 0$. Donde si $\mathbf{v}$ es un vector propio de (2.10), entonces

$$\mathbf{a} = \mathbf{E}^{-1/2}\mathbf{v} \tag{2.11}$$

es un vector propio (2.9) con el mismo valor propio.

Se pueden definir varios test estadísticos basados en dichas matrices y valores propios.

- Traza de Lawley-Hotteling :

$$T = traza(\mathbf{H}\mathbf{E}^{-1}) = \sum_i \lambda_i$$

  siendo λ_i los valores propios no negativos de (2.8).

- Lambda de Wilks

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|} = \left|\mathbf{E}(\mathbf{H} + \mathbf{E})^{-1}\right| = \prod_i \lambda_i$$

- Estadístico de Pillai

$$V = traza(\mathbf{E}(\mathbf{H} + \mathbf{E})^{-1}) = \sum_i \frac{\lambda_i}{1 + \lambda_i}$$

Los estadísticos definidos anteriormente, con la raíz característica ya citada, pueden ser aproximados de forma asintótica a través de una F de Snedecor. Son los estadísticos más utilizados en la mayor parte de los paquetes de software estadístico ya elaborados.

2.3. MANOVA con un factor de variación

Existe un caso particular del MGLM en el que las variables predictoras son categóricas, este caso es denominado Análisis Multivariante de la Varianza (MANOVA). La variable categórica puede indicar el grupo al que pertenece un individuo o el tratamiento aplicado en el diseño del experimento. El objetivo de esta técnica es buscar si existen diferencias entre los grupos o tratamientos.

Dado un número de categorías, grupos o niveles del factor K, las I filas de la matriz de datos $\mathbf{Y}$ estarán divididas en K grupos en los que encontramos I_k individuos en cada uno de ellos, siendo k el número de grupo y cumpliendo que $I = I_1 + I_2 + \ldots + I_K$.

Es conocido que pueden introducirse variables categóricas en un modelo de regresión mediante variables ficticias o indicadores que toman valores 0 y 1 dependiendo de si el individuo está o no en la categoría de referencia. Puede definirse una de estas variables para cada una de las categorías, aunque no podemos incluirlas todas como variables independientes en un modelo de regresión con constante, ya que serían linealmente dependientes y el modelo no podría estimarse debido a que la matriz $\mathbf{X}^T\mathbf{X}$ sería singular.

2.3.1. Matriz de diseño y estimación de los parámetros

Se define $\mathbf{X}$ de tamaño IxK, como la matriz de diseño que contiene las variables predictoras, las columnas contienen los indicadores de todas las categorías menos una junto a una columna de unos que corresponde a la constante

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{I_1} & \mathbf{1}_{I_1} & \mathbf{0}_{I_1} & \dots & \mathbf{0}_{I_1} \\ \mathbf{1}_{I_2} & \mathbf{0}_{I_2} & \mathbf{1}_{I_2} & \dots & \mathbf{0}_{I_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{1}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} & \dots & \mathbf{1}_{I_{K-1}} \\ \mathbf{1}_{I_K} & \mathbf{0}_{I_K} & \mathbf{0}_{I_K} & \dots & \mathbf{0}_{I_K} \end{pmatrix} \tag{2.12}$$

donde $\mathbf{1}_{I_k}$ y $\mathbf{0}_{I_k}$ son los vectores de unos y ceros con I_k elementos.
Otra posible matriz de diseño es aquella que no suprime ninguna de las categorías e incluye la columna de unos.

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{I_1} & \mathbf{1}_{I_1} & \mathbf{0}_{I_1} & \dots & \mathbf{0}_{I_1} & \mathbf{0}_{I_1} \\ \mathbf{1}_{I_2} & \mathbf{0}_{I_2} & \mathbf{1}_{I_2} & \dots & \mathbf{0}_{I_2} & \mathbf{0}_{I_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{1}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} & \dots & \mathbf{1}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} \\ \mathbf{1}_{I_K} & \mathbf{0}_{I_K} & \mathbf{0}_{I_K} & \dots & \mathbf{0}_{I_K} & \mathbf{1}_{I_K} \end{pmatrix} \tag{2.13}$$

En el caso de la ecuación (2.13) no es posible calcular los estimadores de los parámetros ya que el producto por su transpuesta ($\mathbf{X^TX}$) es una matriz singular. Para evitar este problema es posible utilizar, en lugar de la inversa $(\mathbf{X^TX})^{-1}$, una inversa generalizada $(\mathbf{X^TX})^{-}$.

Existen diferentes formas para construir la matriz de diseño $\mathbf{X}$ dando lugar a diferentes parametrizaciones del modelo. En este trabajo resulta de interés utilizar la más sencilla, será cons-

truida incluyendo en la matriz únicamente los indicadores de todas las categorías

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{I_1} & \mathbf{0}_{I_1} & \dots & \mathbf{0}_{I_1} & \mathbf{0}_{I_1} \\ \mathbf{0}_{I_2} & \mathbf{1}_{I_2} & \dots & \mathbf{0}_{I_2} & \mathbf{0}_{I_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} & \dots & \mathbf{1}_{I_{K-1}} & \mathbf{0}_{I_{K-1}} \\ \mathbf{0}_{I_K} & \mathbf{0}_{I_K} & \dots & \mathbf{0}_{I_K} & \mathbf{1}_{I_K} \end{pmatrix} \tag{2.14}$$

Para esta matriz de diseño, considerando que la matriz de contrastes y la matriz de combinaciones de las variables son igual a la identidad, entonces

$$\mathbf{X}^{\mathbf{T}}\mathbf{X} = diag(I_1, I_2, \dots, I_K) = \mathbf{D}_K$$

, es decir la matriz diagonal que contienen el tamaño muestral de los grupos.

También se puede afirmar que

$$\mathbf{R} = \mathbf{D}_K^{-1} = diag(\frac{1}{I_1}, \dots, \frac{1}{I_K})$$

$$\hat{\Omega} = \hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\ \mathbf{Y} = \bar{\mathbf{Y}}$$

$$\mathbf{E} = \mathbf{Y}^T\mathbf{Y} - \bar{\mathbf{Y}}^T\mathbf{D}_K\bar{\mathbf{Y}}$$

y

$$\mathbf{H} = \bar{\mathbf{Y}}^T\mathbf{D}_K\bar{\mathbf{Y}}$$

donde $\bar{\mathbf{Y}}$ es una matriz de tamaño $K \times J$ que contiene los vectores de medias de cada grupo, $\mathbf{E}$ es la matriz con las sumas de cuadrados y los productos escalares dentro de los grupos y $\mathbf{H}$ contiene las sumas de cuadrados y los productos escalares entre los grupos.

Es bien conocido que las variables canónicas o coordenadas discriminantes corresponden a la combinación lineal de las variables observables con mayor poder discriminante, esta combinación se puede obtener calculando los vectores propios de la matriz $\mathbf{E}^{-1}\mathbf{H}$.

La primera coordenada, que genera la $\mathbf{F}$ univariante más grande, puede ser interpretada como aquella que recoge la máxima variabilidad entre los grupos con respecto a la variabilidad dentro de estos. De forma análoga ocurre con las demás variables canónicas con la variabilidad no explicada en las anteriores. El número de coordenadas discriminante corresponde con el número de grupos menos 1, de forma teórica corresponde al mín$(I - K,\ K - 1)$.

Podría afirmarse que es similar a un Análisis de Componentes Principales de las medias teniendo en cuenta la variabilidad existente dentro de los grupos. Por tanto, será posible proyectar las medias de los grupos sobre el espacio de las coordenadas discriminantes. Es posible calcular las coordenadas canónicas de las medias $\bar{\mathbf{Z}}$ mediante los coeficientes de la ecuación (2.11)

$$\bar{\mathbf{Z}} = \bar{\mathbf{Y}}\mathbf{a} = \bar{\mathbf{Y}}\mathbf{E}^{-1/2}\mathbf{v}$$

Igual que se proyectan las medias, es posible hacer una proyección de los individuos originales que permita observar la separación de los grupos. Las coordenadas de los individuos en el espacio canónico serían

$$\mathbf{Z} = \mathbf{Y}\mathbf{a} = \mathbf{Y}\mathbf{E}^{-1/2}\mathbf{v}$$

También es posible sustituir la matriz de productos cruzados $\mathbf{E}$ por la matriz de covarianzas dentro de los grupos

$$\mathbf{S} = \frac{1}{I-K}\mathbf{E}$$

que al diferenciarse únicamente en una constante, la representación resultante será exactamente la misma.

Una propiedad a tener en cuenta en la representación canónica es que, la distancia de Mahalanobis entre las medias en el espacio de inicial, es igual a la distancia euclídea en el espacio de la representación.

La interpretación de las variables canónicas se realiza calculando las correlaciones entre las variables observadas y las canónicas, es decir, de igual forma que en el Análisis Factorial (AF). Estas correlaciones pueden no estar optimizadas y, por lo tanto, tomar valores pequeños.

2.4. Las matrices de combinaciones lineales

Anteriormente hemos mencionado las matrices de combinaciones lineales $\mathbf{C}$ de las medias que nos permite contrastar las columnas de $\mathbf{X}$, si utilizamos la identidad, cada fila corresponde con una variable. En general la matriz contiene S filas

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_S^T \end{pmatrix} \tag{2.15}$$

donde $\mathbf{c}_s^T = (c_{s1}, \ldots, c_{sL})$ contiene coeficientes para contrastar combinaciones lineales de las columnas de $\mathbf{X}$. En el caso del MANOVA con un factor de variación, las combinaciones lineales pueden ser contrastes sobre las medias de los grupos. Se dice que una combinación lineal de las medias con coeficientes $\mathbf{c} = (c_1, \ldots, c_L)^T$ es un contraste si $\sum_{l=1}^{L} c_l = 0$. Por ejemplo, las comparaciones de los grupos por parejas se pueden hacer mediante la matriz de contrastes.

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & -1 & 0 & \ldots & 0 & 0 \\ 1 & 0 & 0 & -1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 0 & 0 & \ldots & -1 & 0 \\ 1 & 0 & 0 & 0 & \ldots & 0 & -1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 1 & 0 & 0 & \ldots & 0 & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & -1 \end{pmatrix} \tag{2.16}$$

Podemos realizar el contraste simultáneo de todas las comparaciones o hacer cada una de ellas por separado. De la misma forma que en los análisis univariantes, se pueden utilizar correcciones para comparaciones múltiples.

En caso de querer introducir combinaciones lineales de las variables, se incluiría en el modelo la matriz $\mathbf{M}$

$$\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_R)$$

para la que $\mathbf{m}_r = (m_{r1}, \ldots, m_{rJ})^T$.

Generalmente se realiza el contraste de todas las variables de forma simultánea, por tanto esta matriz corresponde con la identidad ($\mathbf{M} = \mathbf{I}$), aunque existen casos en los que es de gran utilidad la comparación entre variables, por ejemplo cuando se realiza un Análisis de Perfiles, para los que se calculan las diferencias entre diferentes mediciones de la misma variable, contiguas en el

tiempo, en lugar de las variables originales. La matriz **M** en este caso sería

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & -1 & 1 & \dots & 0 & 0 \\ 0 & 0 & 0 & -1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 \end{pmatrix} \tag{2.17}$$

Estas dos matrices permiten personalizar el modelo a los datos hasta ajustarlo lo más posible al interés del investigador. Es posible combinar ambas matrices para obtener modelos más complejos.

Es bien conocido que las sumas de cuadrados se pueden segmentar en tantos grupos como grados de libertad estén asociados a la hipótesis de partida. Esta separación se puede obtener mediante contrastes ortogonales.

Se dice que dos contrastes $\mathbf{c} = (c_1, \dots, c_L)^T$ y $\mathbf{d} = (d_1, \dots, d_L)^T$ son ortogonales cuando la suma de productos de sus cocientes es 0, $\sum_l c_l d_l = 0$.

Para obtener este tipo de contrastes existe una posible forma conocida como los contrastes de Helmert, este método compara, de forma recursiva, cada grupo con todos los siguientes (o los anteriores). Un posible ejemplo para un caso con cuatro grupos sería

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \mathbf{c}_3 \end{pmatrix} = \begin{pmatrix} 3 & -1 & -1 & -1 \\ 0 & 2 & -1 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \tag{2.18}$$

2.5. Diseños más complejos

En un gran número de diseños se establecen más de dos factores de variación, además puede ser de interés introducir interacciones entre dichos factores, estas opciones las recoge la matriz de contrastes **C**, ya que permite aislar tanto los efectos principales como sus interacciones.

Podemos construir un ejemplo con dos factores de variación que tengan L_1 y L_2 niveles de variación respectivamente. Podemos calcular los grados de libertad para los efectos de cada uno

de ellos como $(L_1 - 1)$ y $(L_2 - 1)$, los grados de libertar para la interacción serían calculados como $(L_1 - 1) \times (L_2 - 1)$. La matriz de contrastes será construida a partir de contrastes ortogonales para contrastar cada uno de los efectos $\mathbf{C}_1$ y $\mathbf{C}_2$ de la forma 2.18, que tendrán tamaños $(L_1 - 1) \times K$ y $(L_2 - 1) \times K$ respectivamente, y la matriz de contrastes de la interacción entre ambos factores $\mathbf{C}_{12}$, con tamaño $((L_1 - 1) \times (L_2 - 1)) \times K$, que resulta de multiplicar cada una de las filas de $\mathbf{C}_1$ por cada una de las de $\mathbf{C}_2$. La matriz de contrastes

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \\ \mathbf{C}_{12} \end{pmatrix} \tag{2.19}$$

Será posible hacer una proyección de los contrastes sobre la representación canónica de las medias para ayudar en la interpretación de los resultados.

Capítulo 3

MANOVA basado en distancias

3.1. Introducción

Existen varios requisitos que deben cumplirse para la correcta aplicación del MANOVA,

- Los datos deben seguir una distribución Normal Multivariante.
- Las matrices de varianzas y covarianzas deben ser homogéneas dentro de los grupos.
- El número de individuos debe ser menor al número de variables.

En la aplicación práctica del MANOVA, en muchas ocasiones, no es posible mantener la condición de Normalidad Multivariante (Xu and Cui, 2008), por ejemplo en los estudios ecológicos que emplean las abundancias como datos, este tipo de estudios presentan una gran cantidad de ceros en sus datos y, generalmente, distribuciones asimétricas.
La razón para que el número de variables deba ser bastante mayor que el de individuos es que, de no ser así, la estimación de los parámetros deja de tener un único valor y pasaría a tener infinitos posibles valores. Sin embargo, en la práctica se dan casos en los que esta condición no se cumple, como es el caso de la genómica, que mide en un reducido número de individuos la expresión de miles de genes.

Es necesario extender la aplicación del MANOVA a aquellas bases de datos que no cumplen dichas condiciones. Existen algunas técnicas alternativas creadas con este fin, por ejemplo, Anderson (2001) que propone un test no paramétrico que se basa en distancias y utiliza las permutaciones para aproximar un estadístico de contraste. Esta técnica recibe por nombre PERMANOVA

y tiene una gran utilidad, sobre todo en datos de ecología. Una de las primeras aplicaciones realizada se puede encontrar en el artículo de McArdle and Anderson (2001), en el que además se realiza una extensión de la técnica para poderlo aplicar a cualquier Modelo Lineal General Multivariante.
Otros ejemplos de técnicas no paramétricas desarrolladas como alternativa al MANOVA se basan en la comparación de los centros de varios grupos, su base se encuentra en una técnica desarrollada por Clarke (1993) y que recibe el nombre de ANOSIM, para su realización podemos encontrar un software denominado PEIMER–E que incluye un módulo para la técnica anterior.

En este trabajo se describirá resumidamente el PERMANOVA, se propondrá la utilización de una técnica similar al esta, pero en el proceso emplea técnicas bootstrap no paramétricas en vez de análisis de permutaciones.

Los pasos para la realización de esta técnica son:

- Calcular la matriz de distancias entre individuos.
- Calcular la pseudo-F como la suma de cuadrados de las distancias *entre* y *dentro* de los grupos.
- Estimar la distribución del estadístico bajo la hipótesis nula.
- Calcular el correspondiente $p - valor$.

En este capítulo se desarrollan los pasos citados anteriormente para realizar los cálculos necesarios. En la sección 3.2 se desarrollará como calcular las distancias entre los individuos para los diferentes tipos de datos que nos podemos encontrar. En el apartado 3.3 encontraremos una explicación un poco más detallada de la técnica PERMANOVA. La siguiente sección (sección **??**) contiene el desarrollo de la nueva técnica, denominada BOOTMANOVA, el apartado 3.4.2 contiene el diseño más sencillo del modelo, que se generaliza en la sección 3.4.3 para el resto de casos más complejos.

3.2. Cálculo de distancias

En este apartado se desarrollarán algunas de las distancias y medidas de similitud que se puedan emplear en la aplicación de las técnicas, siempre que se adapten a los datos. Pueden ser seleccionada cualquiera de ellas, pero, por regla general, se elige en función del ámbito al que pertenece el conjunto de datos sometidos a estudio. Existen medidas de disimilitud, similitud, proximidad o distancia tanto para datos cualitativos como para datos cuantitativos.
En Cuadras Avellanas (1988); Goshtasby (2012); Gray and Markel (1976); Han et al. (2011); Zhang and Srihari (2003) se pueden encontrar resumidos los índices desarrollados en este capítulo.

3.2.1. Distancias en datos continuos

Dada una matriz de datos $\mathbf{Y}_{(IxJ)}$ con I individuos y J variables continuas. Se puede reflejar la matriz de datos como una concatenación de vectores de individuos que contiene la medida de cada individuo en cada una de las J variables observadas.

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_i^T \\ \vdots \\ \mathbf{y}_I^T \end{pmatrix} \tag{3.1}$$

con $\mathbf{y}_i^T = (y_{i1}, \dots, y_{iJ})^T$.

Algunas de las medidas de distancias que se pueden utilizar para datos continuos los podemos encontrar en la siguiente lista:

Euclídea ordinaria :

$$\delta_{il} = \sqrt{\sum_{j=1}^{J} (y_{ij} - y_{lj})^2}$$

Minkowsky :

$$\delta_{il}^r = \left(\sum_{j=1}^{J} |y_{ij} - y_{lj}|^r \right)^{1/r}$$

Camberra :

$$\delta_{il}^{CA} = \sum_{j=1}^{J} \frac{|y_{ij} - y_{lj}|}{y_{ij} + y_{lj}}$$

Bray-Curtis :

$$\delta_{il}^{BC} = \frac{\sum_{j=1}^{J} |y_{ij} - y_{lj}|}{\sum_{j=1}^{J} (y_{ij} + y_{lj})}$$

Divergencia :

$$\delta_{il}^{D} = (\frac{1}{J} \sum_{j=1}^{J} (\frac{|y_{ij} - y_{lj}|}{y_{ij} + y_{lj}})^2)^{1/2}$$

3.2.2. Distancias en datos binarios

Para los datos binarios se calcula una medida de similitud (s_{il}) que posteriormente será convertida a la una medida de distancia utilizando la fórmula $\delta_{il} = \sqrt{1 - s_{il}}$.

Para poder definir algunas de las medidas de similitud, se emplearán dos vectores definidos con datos binarios, y_i y y_j, codificado en función de si existe o no existe la presencia de cada carácter o variable (1 = presencia, 0 = ausencia). Empleando este criterio es posible definir el cuadro 3.1.

i/l	Presente	Ausente	Total
Presente	$a = \mathbf{y}_i^T \mathbf{y}_l$	$b = \mathbf{y}_i^T (1 - \mathbf{y}_l)$	$a + b$
Ausente	$c = (1 - \mathbf{y}_i)^T \mathbf{y}_l$	$d = (1 - \mathbf{y}_i)^T (1 - \mathbf{y}_l)$	$c + d$
Total	a+c	b+d	J=a+b+c+d

Cuadro 3.1: Tabla de contingencia para el cruce de dos individuos

Así quedan definidas a, b, c y d, donde a es el número de individuos que está presente en las dos variables sometidas a estudio, b y c serán aquellas que estén presentes en uno solo de los caracteres y d el número de individuos que no están presentes en ninguna de las variables. Utilizando estas cuatro frecuencias, calculadas en el cuadro 3.1, será posible obtener los coeficientes de similitud y asociación entre las variables.

Algunos de estos índices de similitud se encuentran recogidos en la siguiente lista:

Jaccard :

$$s_{il} = \frac{a}{a + b + c}$$

Dice :

$$s_{il} = \frac{2a}{2a + b + c}$$

Concordancia simple :

$$S_{il} = \frac{a + d}{a + b + c + d}$$

Rogers y Tanimoto :

$$S_{il} = \frac{a + d}{a + 2b + 2c + d}$$

3.2.3. Distancias en datos de diferentes tipos

Puede ocurrir que las variables sean de diferentes tipos, en este caso será necesario calcular, para cada una de las variables, una similaridad (s_{ilj}) y una ponderación (w_{ilj}), que permiten realizar una media ponderada de la similaridad y obtener una similaridad global.

$$s_{il} = \frac{\sum_{j=1}^{J} s_{ilj} w_{ilj}}{\sum_{j=1}^{J} w_{ilj}} \tag{3.2}$$

Los dos índices varían entre 0 y 1. Las similaridades de cada una de las variables varia en función del tipo de variable que sea.

A continuación se desarrolla el calculo de las similaridades y ponderaciones para cada variable en función del tipo de variable:

Variables Binarias :

Similaridad:

- $s_{ilj} = 1$ cuando coinciden ambas variables.
- $s_{ilj} = 0$ cuando no hay coincidencia entre las variables.

Ponderación:

- $w_{ilj} = 0$ cuando existe ausencia de ambas variables, correspondería a la d del cuadro 3.1.
- $w_{ilj} = 1$ para el resto de los casos.

Variables Nominales :

Similaridad:

- $s_{ilj} = 1$ para coincidencias entre variables.
- $s_{ilj} = 0$ para divergencias sin tener en cuenta el número de categorías.

Ponderación:

- $w_{ilj} = 0$ para datos perdidos.
- $w_{ilj} = 1$ para el resto de los casos.

Variables Cuantitativas :

Similaridad:

$s_{ilj} = 1 - \frac{|x_{ij} - x_{lj}|}{R_j}$ donde R_j es el rango de la j-ésima variable.

Ponderación:

- $w_{ilj} = 0$ para datos perdidos.
- $w_{ilj} = 1$ para el resto de los casos.

Es conveniente recordar que con la ecuación (3.2) se está calculando una medida de similaridad, será necesario convertirla en una medida de distancia ($\delta_{il} = \sqrt{1 - s_{il}}$) antes de comenzar a trabajar.

3.3. PERMANOVA

Como ya ha sido mencionado en la introducción de este capítulo (sección 3.1), el PERMANOVA es una técnica descrita por Anderson (2001); McArdle and Anderson (2001). Consiste en realizar un análisis de permutaciones del MANOVA basado en distancias.

Tras realizar el calculo de la matriz de distancias o disimilitudes se puede obtener, a partir de ella, las sumas de cuadrados total y la suma de cuadrado dentro de los grupos, que por diferencia nos permiten calcular la suma de cuadrados entre los grupos.
Utilizando dichas sumas de cuadrados es posible calcular una F de forma análoga a la calculada en el análisis univariante, es decir, habremos obtenido un MANOVA basado en las distancias de los datos originales.
El siguiente paso será realizar la estimación de la distribución muestral suponiendo que la hipótesis nula es cierta, para ello se utilizan las permutaciones, como no es posible cogerlas todas elegiremos al azar un número elevado de permutaciones. En cada una de las permutaciones se recalculará la F creando la distribución buscada.
Por último obtendremos un $p - valor$ asociado, para ello debe calcularse la proporción de los valores de F obtenidas en las permutaciones son mayores que la F que se había calculado con las

distancias de los datos originales.

Una justificación más teórica de la técnica se puede encontrar en McArdle and Anderson (2001).

Existe un programa diseñado por Anderson (2005) que permite realizar los cálculos de los test pertinentes e incluye la posibilidad de incluir modelos lineales de mayor complejidad.

Una técnica análoga a la descrita en esta sección (3.3) es la descrita por Gower and Krzanowski (1999), su propuesta consiste en un análisis de distancias para datos estructurados multivariantes empleando como contraste un Análisis de Coordenadas Principales de las medias de los grupos sobre las que se proyectan el conjunto completo de individuos, aunque está menos extendido por la falta de software para su realización y por haber sido publicado en una revista con menos difusión entre los investigadores por ser propiamente de Estadística.

3.4. BOOTMANOVA

3.4.1. Introducción

Es posible encontrar diversos artículos en los que se utiliza el bootstrap asociados al MANOVA como son Goodnight and Schwartz (1997); Van Aelst and Willems (2011); Krishnamoorthy and Lu (2010), sin embargo ninguno de ellos emplea un MANOVA basado en distancias.

Generalizando el PERMANOVA resumido en la sección 3.3, será utilizado un método bootstrap en lugar de un análisis de permutaciones para estimar la distribución muestral.

3.4.2. Diseños con un factor de variación

Desarrollo

El objetivo de esta técnica es buscar si existen diferencias significativas entre los K grupos definidos. Para ello debemos plantear la hipótesis estadística asociada a este objetivo de la forma

$$\begin{cases} \mathrm{H}_0 \ : \text{Todos los grupos son iguales} \\ \mathrm{H}_a \ : \text{Existen diferencias entre algunos de los grupos} \end{cases} \tag{3.3}$$

o, si la escribimos en notación matemática, como

$$\begin{cases} \mathrm{H}_0 \ : \mu_1 = \mu_2 = \cdots = \mu_K \\ \mathrm{H}_a \ : \exists i, j, \in (1, 2, \ldots K), \mu_i \neq \mu_j \end{cases} \tag{3.4}$$

donde μ_k es el vector de medias del grupo k, donde $k = 1, 2..., K$.

Igual que en el caso anterior, definimos una matriz de datos $\mathbf{Y}_{(IxJ)}$, donde I corresponde con el número de filas, que se encuentran divididas en K grupos con I_k individuos en cada uno de ellos y que cumplen las condiciones $(k = 1, 2..., K)$, $(I = I_1 + I_2 + ... + I_K)$, y J corresponde con el número de variables de la matriz. Además, utilizaremos la matriz de diseño $\mathbf{X}_{(IxK)}$ con I filas y K columnas construida de la forma descrita en la ecuación 2.14.

Como ya se mencionaba en las introducciones anteriores, en la técnica BOOTMANOVA será necesario construir una matriz de distancias, a través de los cálculos de similitudes o disimilitudes que han sido resumidos en el apartado anterior (3.2).
A continuación se obtendrán las sumas de cuadrados a través de las distancias calculadas, a las que denominaremos $\mathbf{\Delta}_{I\times I} = (\delta_{ij})$ y que contendrá ceros en la diagonal. De forma análoga, la matriz que contiene los cuadrados de las distancias la llamaremos $\mathbf{\Delta}^2 = (\delta_{ij}^2)$.
Empleando dichas distancias podemos realizar los cálculos para obtener las sumas de cuadrados total de la siguiente forma

$$SC_T = \frac{1}{I} \sum_{i=1}^{I-1} \sum_{l=i+1}^{N} \delta_{il}^2 \tag{3.5}$$

Si se tiene en cuenta que la matriz completa es simétrica, es conocido que la suma de los cuadrados de las distancias de cada punto al centroide es igual a la suma de los cuadrados de las interdistancias entre todos los puntos dividido por el número de ellos

$$\sum_{i=1}^{I} d^2(\mathbf{x}_i, \bar{\mathbf{x}}) = \frac{1}{I} \sum_{i<l} d^2(\mathbf{x}_i, \mathbf{x}_l) \tag{3.6}$$

la suma de cuadrados total corresponde con la suma de las distancias de la diagonal inferior de la matriz completa dividida por el número de observaciones, por tanto, es posible escribir la ecuación (3.5) en forma matricial como

$$SC_T = \frac{1}{I}\mathbf{1}^T\frac{1}{2}\mathbf{\Delta}^2\mathbf{1} \tag{3.7}$$

donde $\mathbf{1}_I$ es un vector columna con I unos.

El caso de las sumas de cuadrados dentro de los grupos tiene mayor complejidad, ya que las distancias que debemos calcular serán hasta el centroide de cada uno de los grupos. Utilizaremos de nuevo la propiedad (3.6) que se puede ver ilustrada en la figura 3.1 Además incluiremos una

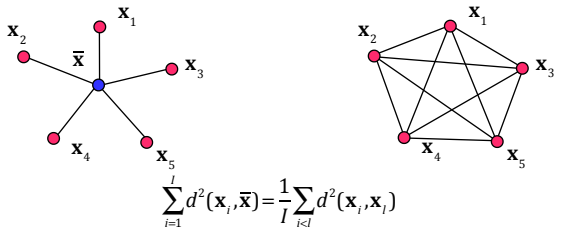

Figura 3.1: Sumas de cuadrados de las distancias

matriz de pertenencia $\mathbf{P} = (\mathbf{p_{il}})$ de tamaño $I \times I$ y que está compuesta por ceros y unos, $p_{il} = 1$ si los individuos i y l pertenecen al mismo grupo y $p_{il} = 0$ si pertenecen a grupos distintos. La matriz diagonal que contiene los tamaños muestrales de los grupos será denominada $\mathbf{D}_K = diag(I_1, \dots, I_K)$. Así es posible definir la matriz $\tilde{\Delta} = \mathbf{P} * \Delta$, donde $*$ es el producto elemento a elemento de las dos matrices, de esta forma si dos individuos i y l pertenecen al mismo grupo $\tilde{\delta}_{il} = \delta_{il}$ y sino $\tilde{\delta}_{il} = 0$.

Una vez definidas las matrices anteriores, es posible calcular las sumas de cuadrados dentro de los grupos empleando la siguiente formula definida en forma matricial

$$SC_D = \mathbf{1}_K^T\mathbf{D}_K^{-1/2}\mathbf{X}^T(\frac{1}{2}\tilde{\Delta}^2)\mathbf{X}\mathbf{D}_K^{-1/2}\mathbf{1}_K \tag{3.8}$$

La suma de cuadrados dentro de los grupos puede definirse como la diferencia entre la suma de cuadrados total y la suma de cuadrados dentro de los grupos

$$SC_E = SC_T - SC_D \tag{3.9}$$

Esta parte del proceso se puede ilustrar a través de la figura 3.2, en el que se recoge el cálculo de las distancias junto a las sumas de cuadrados del modelo.

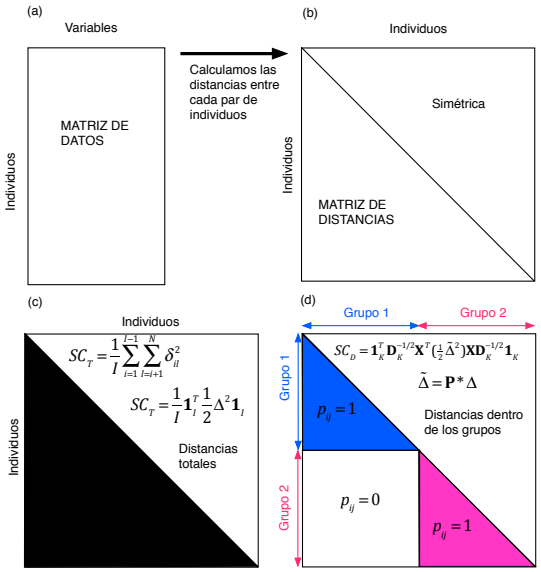

Figura 3.2: Cálculo de las distancias. (a) Matriz de datos brutos. (b) Matriz simétrica que contiene las distancias. (c) Distancias totales. (d) Distancias dentro de los grupos.

De forma análoga al ANOVA, a partir de las sumas de cuadrados se puede obtener un pseudo estadístico F

$$F = \frac{SC_E/(K-1)}{SC_D/(I-K)} \tag{3.10}$$

Por regla general, su distribución no sigue una F de Snedecor ya que no esperamos que las variables sean normales, sin embargo, si tenemos una única variable dependiente, se cumple la condición de normalidad y la distancia utilizada es la distancia euclídea usual, la F calculada coincide con la F univariante del procedimiento tradicional.

Para obtener la distribución muestral del estadístico se utilizará el método bootstrap. Este método se encuentra desarrollado en Efron (1979); Efron and Tibshirani (1986, 1994) y consiste en realizar N veces un remuestreo con reemplazamiento de I elementos de la muestra original, estimando la F para cada caso, que nos permite obtener la distribución muestral. Generalmente N se encuentra entre 1000 y 2000 muestras.

En nuestro caso cogeremos el mismo número de individuos que tenemos en la muestra es decir $t = I$.

Retomamos el planteamiento de la hipótesis nula realizada en la ecuación (3.4), si la hipótesis nula es cierta y los vectores de medias de todos los grupos son iguales, las observaciones serían intercambiables y no influirá haber realizado el remuestreo sobre los individuos, ya que las etiquetas correspondientes a los grupos pueden ser asignadas al azar. Las F obtenidas de cada uno de los remuestreos realizados, a las que denominaremos a partir de ahora F^{π}, y que formarán la distribución muestral del estadístico, nos permitirán, al compararlas con el valor original, obtener un $p - valor$ asociado.

Definimos el $p - valor$ como la probabilidad de que, siendo cierta la hipótesis nula (ecuación (3.4)) los resultados obtenidos con la muestra realizada a través del método bootstrap sean más extremos que el resultado obtenido con la muestra original.

De forma análoga al PERMANOVA descrito por Anderson (2001) podemos definir dos formas de obtener dicho $p - valor$

$$p = \frac{\text{Número de } F^{\pi} \geq F}{\text{Número total de } F^{\pi}} \tag{3.11}$$

0 bien

$$p = \frac{(\text{Número de } F^{\pi} \geq F) + 1}{(\text{Número total de } F^{\pi}) + 1} \tag{3.12}$$

Es posible realizar una justificación teórica del modelo propuesto describiendo la comparación entre grupos a partir del modelo lineal general multivariante en su notación matricial. Para ello partimos de una matriz de datos centrados a la que denominamos $\mathbf{Y}$, en su matriz de covarianzas ($\mathbf{Y^TY}$) y en su matriz de productos escalares ($\mathbf{YY^T}$) se encuentra la información de mayor relevancia para un gran número de técnicas de Análisis Multivariante. Gower (1966) describe como puede obtenerse la matriz de distancias a partir de la de productos escalares, independientemente del tipo de matriz de distancias obtenida.

Utilizaremos la matriz de diseño $\mathbf{X}$ (ecuación (2.14)). Como hemos dicho anteriormente, el método de cálculo tradicional es a través de las sumas de cuadrados. Es posible obtener estas sumas de cuadrados a partir de las matrices de productos curzados $\mathbf{Y^TY}$. Si sumamos los valores de la diagonal de esta matriz, es decir si calculamos su traza, se puede demostrar que se obtiene la Suma de Cuadrados Total

$$SC_T = tr(\mathbf{Y}^T\mathbf{Y}) \tag{3.13}$$

Emplearemos el modelo lineal general multivariante descrito en 2.1 para contrastar el caso particular en el que la hipótesis es

$$\hat{\Omega} = \mathbf{B} = \mathbf{0} \tag{3.14}$$

Utilizando el estimador mínimo cuadrático especificado en la ecuación (2.2) se pueden obtener los vectores estimados $\hat{\mathbf{Y}} = \mathbf{XB} = \hat{\mathbf{H}}\mathbf{Y}$, donde $\hat{\mathbf{H}} = \hat{\mathbf{X}}(\hat{\mathbf{X}}^T\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}^T$. La matriz de residuales puede ser calculada como la diferencia de los datos reales y los estimados $\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \hat{\mathbf{H}})\mathbf{Y}$.
Entonces la matriz de partida, es decir, la matriz de sumas de cuadrados y productos, puede descomponerse en una parte que está explicada por el modelo y una parte residual. En forma matricial el modelo puede ser escrito como

$$\mathbf{Y}^T\mathbf{Y} = \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} + \hat{\mathbf{U}}^T\hat{\mathbf{U}} \tag{3.15}$$

Si la suma de cuadrados total (SC_T) antes la habíamos asociado a la matriz original, podremos descomponerla en una parte explicada, que corresponderá a las sumas de cuadrados entre grupos, y una parte residual, que estará asociada a la variabilidad que existen dentro de los grupos. Estas dos sumas de cuadrados podrán calcularse, de forma análoga al caso anterior, empleando las trazas:

$$SC_E = tr(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}) \tag{3.16}$$

$$SC_R = tr(\hat{\mathbf{U}}^T\hat{\mathbf{U}}) \tag{3.17}$$

Utilizando las definiciones de las sumas de cuadrados ((3.13), (3.16), (3.17)) se pude afirmar que

$$tr(\mathbf{Y}^T\mathbf{Y}) = tr(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}}) + tr(\hat{\mathbf{U}}^T\hat{\mathbf{U}}) \tag{3.18}$$

y por tanto, podríamos definir un pseudo estadístico de contraste apropiado a través de una F calculada a partir de las trazas y los grados de libertad del modelo

$$F = \frac{tr(\hat{\mathbf{Y}}^T\hat{\mathbf{Y}})/K - 1}{tr(\hat{\mathbf{U}}^T\hat{\mathbf{U}})/I - K} \tag{3.19}$$

En este estadístico, a diferencia de la hipótesis general del MLGM (2.3), no se emplean las covarianzas y correlaciones entre las variables o individuos, ya que $\mathbf{C} = \mathbf{I}$ y $\mathbf{M} = \mathbf{I}$. Igual que el estadístico F descrito en la ecuación (3.10), si solo existe una variable estaremos hablando del estadístico utilizado en el ANOVA.

A continuación será necesario estimar la distribución muestral, para ello utilizaremos el bootstrap no paramétrico sobre los individuos.
Dada la matriz de distancias a la que denominamos $\mathbf{\Delta} = (\delta_1, \delta_2, \dots, \delta_I)$, siendo δ_i el vector de distancias del elemento $i = (1, 2, \dots, I)$ tomaremos una muestra elegida al azar con I elementos a la que denominaremos $\mathbf{\Delta}^* = (\delta_1^*, \delta_2^*, \dots, \delta_I^*)$. Recalcularemos el valor del estadístico (F) a partir de la matriz $\mathbf{\Delta}^*$, valores a los que denominaremos F^π como en la explicación anterior. Repetiremos este proceso un número N de veces fijado previamente, antes de comenzar el análisis.

Por último se calculará el p-valor asociado al estadístico de la misma forma que anteriormente, empleando la fórmula (3.11) o (3.12).

Si la matriz de partida es la matriz de productos escalares entre individuos, puede obtenerse la misma descomposición teniendo en cuenta que dos matrices **A** y **B** cumplen que $tr(\mathbf{AB}) = \mathbf{tr(BA)}$. Por tanto si $tr(\mathbf{Y^TY}) = \mathbf{tr(YY^T)}$, se puede conseguir la misma división del modelo que en (3.18), pero utilizando las matrices de productos escalares

$$tr(\mathbf{YY}^T) = tr(\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T) + tr(\hat{\mathbf{U}}\hat{\mathbf{U}}^T) \tag{3.20}$$

En este caso la partición sigue siendo posible, aunque no conozcamos **Y**, a partir de la matriz de productos escalares ($\mathbf{YY^T}$) ya que se pueden obtener las estimaciones como

$$\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T = \hat{\mathbf{H}}(\mathbf{YY}^T)\hat{\mathbf{H}}$$

$$\hat{\mathbf{U}}\hat{\mathbf{U}}^T = (\mathbf{I} - \hat{\mathbf{H}})(\mathbf{YY}^T)(\mathbf{I} - \hat{\mathbf{H}})$$

Como ya habíamos citado anteriormente, la matriz de productos escalares puede calcularse a través de cualquier matriz de distancias observadas $\mathbf{\Delta} = (\delta_{ij})$ utilizando de la formula descrita por Gower (1966):

$$\mathbf{G} = (\mathbf{I} - \tfrac{1}{I}\mathbf{1}_I\mathbf{1}_I^T)(\tfrac{1}{2}\mathbf{\Delta}^2)(\mathbf{I} - \tfrac{1}{I}\mathbf{1}_I\mathbf{1}_I^T) \tag{3.21}$$

siendo **G** la matriz de productos escalares que queremos calcular.
Dicha matriz **G** puede descomponerse como se ha indicado en la ecuación (3.20) y, por lo tanto, calcular la suma de cuadrados total a partir de ella $SC_T = tr(\mathbf{G})$. Siguiendo el proceso descrito para el caso de las matrices de sumas de cuadrados y productos se obtendría el estadístico de contraste como

$$F = \frac{tr\left(\hat{\mathbf{H}}\mathbf{G}\hat{\mathbf{H}}\right)/_{K-1}}{tr\left[\left(\mathbf{I} - \hat{\mathbf{H}}\right)\mathbf{G}\left(\mathbf{I} - \hat{\mathbf{H}}\right)\right]/_{I-K}} \tag{3.22}$$

que, empleando de nuevo las fórmulas (3.11) o (3.12), puede obtenerse un p-valor asociado a dicho estadístico.

3.4.3. Diseño generalizado

La hipótesis contrastada en este caso debe incluir la matriz de contrastes **C** que permita aislar los efectos

$$\mathbf{\Omega} = \mathbf{CB} = \mathbf{0} \tag{3.23}$$

En este contexto no nos interesa contrastar las variables, por ello de la hipótesis del MLGM (2.3) hemos suprimido **M**.

Se utilizarán las ecuaciones 2.4, 2.5 y 2.6 para contrastar dicha hipótesis con estadísticos relacionados con las raíces características de $\mathbf{HE}^{-1}$.

El estimador de $\mathbf{\Omega}$ es ahora

$$\hat{\mathbf{\Omega}} = \mathbf{C}\hat{\mathbf{B}} = \mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.24}$$

Para contrastar esta hipótesis es posible utilizar un pseudo estadístico F que tenga en cuenta los grados de libertad de $\mathbf{C}\hat{\mathbf{B}}$, $(S-1)$, de la forma

$$F = \frac{tr(\hat{\mathbf{\Omega}}^T\mathbf{R}^{-1}\hat{\mathbf{\Omega}})/(S-1)}{tr(\hat{\mathbf{U}}^T\hat{\mathbf{U}})/(I-K)} \tag{3.25}$$

El denominador del estadístico de contraste F descrito en la ecuación (3.19) no depende del planteamiento de la hipótesis por lo que se mantiene igual en este caso.

Este estadístico es igual al cuadrado de una t de *Studen* con $(I-K)$ grados de libertad si se trata de un solo contraste, una única variable y se ha usado la distancia euclídea, o lo que es lo mismo, sigue una F de *Snedecor* con 1 y $(I-K)$ grados de libertad.

Al incluir la matriz **C** es posible comparar diseños más complejos, aislar los efectos o realizar las comparaciones por parejas.

Igual que en el caso más sencillo de la hipótesis (ecuación (3.14)), es posible utilizar la matriz de productos escalares, **G**, para contrastar la hipótesis (3.24) teniendo en cuenta que

$$tr(\hat{\mathbf{\Omega}}^T\mathbf{R}^{-1}\hat{\mathbf{\Omega}}) = tr(\hat{\mathbf{\Omega}}^T\mathbf{R}^{-1/2}\mathbf{R}^{-1/2}\hat{\mathbf{\Omega}}) = tr(\mathbf{R}^{-1/2}\hat{\mathbf{\Omega}}\hat{\mathbf{\Omega}}^T\mathbf{R}^{-1/2}) = tr(\mathbf{R}^{-1/2}\mathbf{C}\hat{\mathbf{B}}\hat{\mathbf{B}}^T\mathbf{C}^T\mathbf{R}^{-1/2})$$

$$= tr(\mathbf{R}^{-1/2}\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\mathbf{R}^{-1/2})$$

$$= tr(\mathbf{R}^{-1/2}\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{G}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\mathbf{R}^{-1/2})$$

La pseudo F obtenida para este caso sería

$$F = \frac{tr\left(\mathbf{R}^{-1/2}\mathbf{C}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{G}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{C}^T\mathbf{R}^{-1/2}\right)\Big/(S-1)}{tr\left[\left(\mathbf{I}-\hat{\mathbf{H}}\right)\mathbf{G}\left(\mathbf{I}-\hat{\mathbf{H}}\right)\right]/(I-K)} \tag{3.26}$$

Tanto en el caso del estadístico (3.25) como en el del 3.26 se repetiría el remuestreo de los individuos para obtener N muestras con una F^{π} asociada a cada una de ellas, utilizando una de las dos ecuaciones ((3.11) o (3.12)), igual que en los casos anteriores, se calculará el pvalor asociado al contraste.

Capítulo 4

Representaciones Gráficas

Es conocida, y de gran interés, la representación gráfica de los MLGM, y más concretamente del MANOVA, que permita el estudio de la dimensionalidad asociada a la hipótesis alternativa. Esta representación recibe el nombre de *Análisis Canónico de Poblaciones* o *Coordenadas Discriminantes*. El objetivo de este apartado es buscar una representación gráfica que se pueda asociar a los MANOVAs basados en distancias que permitan realizar un mejor estudio de la hipótesis alternativa de estos modelos.

El Análisis de Coordenadas Principales utiliza la matriz de productos escalares $\mathbf{G}$ como base, por tanto, puede ser de utilidad para la representación gráfica de cualquiera de los MANOVAs basados en distancias que se están estudiando en este trabajo.

4.1. Análisis de Coordenadas Principales

Obtenida la matriz de productos escalares $\mathbf{G}$ a partir de una matriz de distancias observadas $\mathbf{\Delta}$ empleando la ecuación (3.21), se buscar encontrar la configuración de puntos $\mathbf{Z}$ que reproduzca lo más fielmente posible los productos escalares y, por añadidura, las distancias entre individuos. La técnica que realiza este proceso es conocida como *A*nálisis de Coordenadas Principales, propuesta por Gower (1966).

La descomposición en valores y vectores propios de la matriz $\mathbf{G}$ podemos escribirla como

$$\mathbf{G} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}} \tag{4.1}$$

donde $\mathbf{V}$ es la matriz que contiene los vectores propios de $\mathbf{G}$ y $\mathbf{\Lambda}$ la matriz diagonal que encierra los valores propios ordenados de mayor a menor, $\mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \lambda_2, \dots, \lambda_I)$.
Se cumple que si la matriz de vectores propios es definida positiva, todos los valores propios son no negativos y, por tanto existe una configuración euclídea que nos permite reproducir las distancias observadas en una dimensión $(I - 1)$.

Las coordenadas buscadas se encuentran en la matriz

$$\mathbf{Z} = \mathbf{V}\mathbf{\Lambda}^{1/2} \tag{4.2}$$

Las columnas de $\mathbf{Z}$ contienen la variabilidad del modelo que, al igual que los valores propios, se ordenan de forma decreciente, lo que permite elegir las primeras columnas para realizar una representación en dimensión reducida. Es posible calcular la cantidad de variabilidad que se encuentra recogida en las T primeras coordenadas utilizando los valores propios de la siguiente forma

$$\frac{\sum_{j=1}^{T} \lambda_j}{\sum_{j=1}^{I-1} \lambda_j}$$

Estas coordenadas coincidirán con las de las Componentes Principales si las distancia utilizada es la euclídea. Este proceso puede hacerse con cualquier tipo de distancia aunque, si esta difiere de la euclídea, pueden existir valores negativos y de pequeña magnitud. En este caso utilizaremos la opción más sencilla para eliminar este problema, reconstruiremos la matriz $\mathbf{G}$ sustituyendo los valores negativos por ceros, es decir la aproximaríamos a la matriz semidefinida positiva más cercana. Esta es la solución más sencilla, sin embargo en la literatura se pueden encontrar múltiples alternativas para este problema.

La utilización de esta técnica (ACoP) tras realizar cualquiera de las dos técnicas descritas con el MANOVA basado en distancias, correspondería con la aplicación de un Análisis de Componentes Principales (ACP) una vez realizado el MANOVA convencional con una pequeña diferencia, en el ACP se busca encontrar las direcciones que hacen que la variabilidad total sea máxima aunque la variabilidad *entre grupos* no lo sea, para encontrar estas últimas deberemos sustituirlo por un Análisis Canónico que permite maximizar la variabilidad *entre grupos* en relación con la variabilidad *dentro* de estos.

Para este caso, la propuesta realizada por Gower and Krzanowski (1999) fue realizar un ACoP sobre los centroides que recoge mejor que el ACoP las diferencias entre los grupos, aunque no la tenga en cuenta. Con este mismo objetivo Anderson and Willis (2003) propone realizar el Análisis Canónico sobre las Coordenadas Principales.

4.2. Coordenadas Principales de la matriz de medias

Para realizar este análisis debemos calcular la matriz de distancias entre los centroides, a esta matriz la denominaremos $\bar{\mathbf{\Delta}}$. Es posible calcular dicha matriz como describieron Gower and Krzanowski (1999) a partir de la matriz de distancias original, $\mathbf{\Delta}$, la matriz diagonal que contiene los tamaños muestrales de los grupos, $\mathbf{D_K}$, y la matriz de diseño con los indicadores utilizados en los apartados anteriores, $\mathbf{X}$

$$\bar{\mathbf{\Delta}} = \mathbf{D_K^{-1}X^T\Delta XD_K^{-1}} \tag{4.3}$$

Una vez obtenida la matriz de distancias a los centroides, realizaremos un ACoP utilizando dicha matriz. Para ello seguiremos el proceso descrito anteriormente, comenzaremos realizando los cálculos pertinentes para obtener la matriz de productos escalares a partir de esta matriz de distancias

$$\bar{\mathbf{G}} = (\mathbf{I} - \tfrac{1}{K}\mathbf{1}_K\mathbf{1}_K^T)(\tfrac{1}{2}\bar{\mathbf{\Delta}}^2)(\mathbf{I} - \tfrac{1}{K}\mathbf{1}_K\mathbf{1}_K^T) \tag{4.4}$$

A continuación debemos obtener una descomposición en valores y vectores propios, en este caso la matriz que contiene los vectores propios será $\bar{\mathbf{V}}$ y $\bar{\mathbf{\Lambda}}$ la matriz diagonal con los valores propios que, igual que en el caso anterior, estarán ordenados de forma descendiente

$$\bar{\mathbf{G}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}}\bar{\mathbf{V}}^T \tag{4.5}$$

Por último, debemos realizar los cálculos que nos permiten obtener las Coordenadas Principales

$$\bar{\mathbf{Z}} = \bar{\mathbf{V}}\bar{\mathbf{\Lambda}}^{1/2} \tag{4.6}$$

Es posible tener en cuenta los tamaños muestrales de los grupos si se utiliza la versión ponderada descrita por los mismos autores, pero no serán explicados en este documento.

Para realizar la representación de los individuos iniciales sobre el gráfico realizado con las medias, es posible utilizar una fórmula que fue propuesta por Gower (1968).

Suponemos que podemos calcular la matriz de distancias de los puntos de la matriz original $\mathbf{Y}$ a las medias, a dicha matriz la denominaremos $\mathbf{\Delta}_{\bar{\mathbf{Y}}}$ y tendrá dimensión IxK ya que contiene las distancias de los I individuos a las centroides de los K grupos. A partir de las matrices descritas hasta ahora en esta sección 4.2 y con el vector $\bar{\mathbf{g}}$ que contiene los elementos de la diagonal de $\bar{\mathbf{G}}$, podemos obtener las coordenadas de los individuos en la representación de las medias, que denominaremos $\mathbf{Z}_{\bar{\mathbf{Z}}}$

$$\mathbf{Z}_{\bar{\mathbf{Z}}} = \tfrac{1}{2}(\mathbf{1}_I \bar{\mathbf{g}} - \Delta_{\bar{\mathbf{Y}}})\bar{\mathbf{Z}}\bar{\Lambda}^{-2} \tag{4.7}$$

Solo cuando tenemos datos continuos podemos calcular los centroides y las distancias de los individuos a dichos puntos. Cuando los datos son binarios los centroides no tienen porqué ser un vector de datos binarios, lo que imposibilitaría el cálculo de las distancias a los centroides que permiten la representación de los individuos sobre el gráfico creado a partir de las medias. Se puede plantear como posible solución para esta problemática podría proponerse calcular las Coordenadas Principales en la dimensión completa, a continuación obtener los centros y, a partir de ellos, las distancias.

4.3. Regiones de confianza bootstrap para los centroides

Construiremos ahora regiones de confianza para los centroides similares a las que se utilizan en el Análisis Canónico, pero ahora basadas en bootstrap. Se trata de perturbar los datos para mostrar la variabilidad de los centroides y usar los resultados para mostrar la región de confianza. Las regiones se utilizan para comprobar las estructura de la hipótesis alternativa y sirven para hacer contrastes aproximados de comparación de parejas de medias. Un procedimiento basado en bootstrap para la versión clásica del Análisis Canónico puede encontrarse en Duarte et al. (1998). Las regiones resultantes serán similares a las obtenidas en Amaro et al. (2008).

El procedimiento bootstrap será diferente del utilizado en los apartados anteriores ya que entonces el remuestreo lo hacemos bajo la hipótesis de que todos los grupos son iguales mientras que ahora, lo haremos en el supuesto de que hay diferencias entre los grupos. Haremos el muestreo bootstrap dentro de cada grupo.

Dada la matriz de distancias a la que denominamos $\Delta = (\delta_1, \delta_2, \dots, \delta_I)$, siendo δ_i el vector de distancias del elemento $i = (1, 2, \dots, I)$. Vamos a tomar ahora los individuos divididos en K grupos, la matriz de distancias es ahora

$$\Delta = (\delta_{1(1)}, \dots, \delta_{n_1(1)}, \dots, \delta_{1(k)}, \dots, \delta_{n_k(k)}, \dots, \delta_{1(K)}, \dots, \delta_{n_k(K)})$$

Tomamos muestras al azar con reemplazamiento dentro de cada uno de los grupos para obtener

$$\Delta_b^* = (\delta_{1(1)}^*, \dots, \delta_{n_1(1)}^*, \dots, \delta_{1(k)}^*, \dots, \delta_{n_k(k)}^*, \dots, \delta_{1(K)}^*, \dots, \delta_{n_k(K)}^*)$$

con $b = 1, \dots, B$ siendo B el número de muestras bootstrap.

Calculamos las distancias entre las medias como en 4.3,

$$\bar{\Delta}_b^* = \mathbf{D_K^{-1}} \mathbf{X^T} \Delta_b^* \mathbf{X} \mathbf{D_K^{-1}} \tag{4.8}$$

los productos escalares como en 4.4

$$\bar{\mathbf{G}}_b = (\mathbf{I} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T)(\frac{1}{2}(\bar{\Delta}_b^*)^2)(\mathbf{I} - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T)$$

y a partir de la descomposición en valores y vectores propios como en 4.5

$$\bar{\mathbf{G}}_b = \bar{\mathbf{V}}_b \bar{\Lambda}_b \bar{\mathbf{V}}_b^T$$

calculamos las coordenadas principales para las medias como en 4.6

$$\bar{\mathbf{Z}}_b = \bar{\mathbf{V}}_b \bar{\mathbf{\Lambda}}_b^{1/2} \tag{4.9}$$

Cada una de las matrices $\bar{\mathbf{Z}}_b$ contiene las réplicas bootstrap de las coordenadas de las medias de los grupos en la representación. Antes de utilizarlas en la representación gráfica tenemos que tener en cuenta algunos problemas que podemos encontrar debido a que se puede presentar variabilidad adicional por el método para obtener los valores y vectores propios. Las réplicas para las coordenadas de los puntos pueden mostrar variabilidades importantes incluso cuando los valores ajustados son similares. La variabilidad adicional se debe a varias posibles causas

- Una reflexión de los ejes, ya que los signos de los vectores propios son únicos salvo el signo.
- Una inversión en el orden de los valores propios, que puede ocurrir especialmente cuando tienen magnitudes similares.

- Una rotación de los vectores propios aun cuando el espacio bidimensional generado es el mismo.

- Una compresión de los valores debida al remuestreo ya que, al repetirse algunos de los puntos la variabilidad de las remuestras puede ser un poco más pequeña.

El primero de los problemas puede solucionarse simplemente calculando el producto escalar de cada vector propio de la matriz inicial con el correspondiente de la réplica bootstrap y cambiar el signo de este si el producto escalar es negativo. El resto de las situaciones es un poco más difícil, para corregirlas utilizaremos la técnica denominada *Análisis Procrustes* que describimos en la siguiente subsección. Un desarrollo adicional de los problemas que nos podemos encontrar puede verse en Milan and Whittaker (1995).

4.3.1. Procustes

Tenemos dos configuraciones de puntos, la configuración inicial $\bar{\mathbf{Z}}$ para los datos originales obtenida de las coordenadas principales de las medias y la configuración $\bar{\mathbf{Z}}_b$ obtenida para las medias en una réplica bootstrap. Se supone que ambas están centradas de forma que podemos evitar el problema de la traslación en el método Procrustes.

Consideramos que la primera es fija y que transformamos la segunda para conseguir que coincidan ambas tan próximamente como sea posible. Más concretamente se trata de obtener una nueva matriz $\bar{\mathbf{R}}_b = t_b\bar{\mathbf{Z}}_b\mathbf{T}_b$ donde t_b es una constante y $\mathbf{T}_b$ una matriz ortogonal, de forma que la discrepancia entre $\bar{\mathbf{Z}}$ y $\bar{\mathbf{Z}}_b$ sea mínima, es decir, se trata de rotar y re-escalar $\bar{\mathbf{Z}}_b$ hasta que coincida lo máximo posible con $\bar{\mathbf{Z}}$. Si consideramos la matriz $\mathbf{C}_b = \bar{\mathbf{Z}}_b^T\bar{\mathbf{Z}}$ y su descomposición en valores singulares

$$\mathbf{C}_b = \mathbf{P}_b\Delta_b\mathbf{Q}_b^T$$

Entonces t_b y $\mathbf{T}_b$ pueden calcularse como

$$\mathbf{T}_b = \mathbf{P}_b\mathbf{Q}_b^T$$

y

$$t = \frac{tr(\mathbf{T}_b^T\bar{\mathbf{Z}}_b\mathbf{R}_b\bar{\mathbf{Z}})}{tr(\bar{\mathbf{Z}}_b^T\mathbf{R}_b\bar{\mathbf{Z}})}$$

Se supone que, ahora, ambas configuraciones son comparables y pueden representarse en el mismo espacio. Sustituimos, entonces, $\bar{\mathbf{Z}}_b \Leftarrow \bar{\mathbf{R}}_b$ para obtener la réplica de la configuración.

Tenemos entonces un conjunto de réplicas para cada una de las coordenadas de los grupos. Para este conjunto de coordenadas es posible representar una elipse de concentración no paramétrica o una envolvente convexa de los puntos. Podemos calcular la elipse con el procedimiento descrito en De Leeuw and Meulman (1986).

Capítulo 5

Aplicación práctica

A lo largo del capítulo 3 hemos desarrollado la parte teórica de diferentes MANOVAs basados en distancias y una posible representación gráfica para ellos. En este capítulo procederemos a ilustrar dichas técnicas con algunos ejemplos prácticos con datos reales.

Realizaremos la aplicación práctica sobre varias bases de datos diferentes. Las bases de datos, recogidas a través de microarrays, contienen información genética que es medida a través de la intensidad lumínica. Existen varios tipos de microarrays ya que cada una de las empresas crea los suyos propios con una maquinaria específica de trabajo para la lectura y extracción de los datos. Las conclusiones con cualquiera de ellos deben ser análogas. En este trabajo todos los microarrays utilizados pertenecen a Affymetrix.

Antes de comenzar realizaremos una revisión bibliográfica de aplicación de las técnicas desarrolladas en la teoría. Al realizar la revisión bibliográfica sobre el PERMANOVA hemos encontrado que no existen un número elevado de artículos que empleen esta técnica sobre microarrays y, la mayor parte de ellos, están asociados al campo de la ecología, existe un número muy reducido que está asociado a datos genéticos relacionados con alguna enfermedad.
Sobre el BOOTMANOVA no se ha encontrado ninguna aplicación debido a la innovación de la técnica.

Las aplicaciones de estas técnicas serán realizadas en el software R (R Core Team, 2016). Para la aplicación del PERMANOVA existe un paquete ya desarrollado en el software, "vegan" (Oksanen et al., 2017), con varias funciones que permiten realizar el análisis.

Sin embargo, se ha desarrollado un paquete dentro del software que permite calcular ambas funciones de forma matricial y realizar la representación gráfica asociada a estos análisis.
Como hemos desarrollado en el capítulo 3, las dos técnicas siguen una estructura similar, por ello en el paquete permite

- El cálculo de la matriz de distancias.
- Realizar el análisis de la varianza previo.
- Calcular el p-valor asociado al contraste general y a los efectos través de la estimación de la distribución correspondiente para cada una de las técnicas.
- Realizar los contrastes Post Hoc.
- Dibujar la representación gráfica asociada a cada una de las técnicas para la interpretación detallada de los resultados.

Dicho paquete utiliza funciones propias y algunas importadas del paquete "MultiBiplotR" (Vicente-Villardón, 2018). Es posible descargar el paquete de http://biplot.usal.es/classicalbiplot/permanova.

La finalidad de esta aplicación practica es ilustrar la utilidad de los MANOVAS basados en distancias en diferentes campos y con datos que no tienen todos las mismas características.

5.1. Diferenciación de enfermedades mentales

5.1.1. Descripción de los datos

Este conjunto de datos se ha obtenido de un artículo escrito por Iwamoto et al. (2004). El artículo original pretende encontrar genes que diferencien a los enfermos de tres afecciones mentales entre ellos y con individuos sanos utilizando muestras de las cortezas prefrontales de los pacientes una vez que han fallecido. Las muestras han sido tomadas de pacientes que han padecido depresión, trastorno de bipolaridad, esquizofrenia y, para el caso de los controles, que no hayan padecido ninguna de las enfermedades anteriores. Las muestras de cortezas prefrontales utilizadas para este estudio han sido proporcionadas por la Stanley Foundation Brain Collection. Este conjunto de muestras relacionadas está recogido en el Gene Expression Omnibus (GEO),

organizadas en una serie con clave de localización GSE12654.

En dicho artículo (Iwamoto et al., 2004) comparan cada una de las enfermedades con los controles, obteniendo diferencias en una serie de genes, una vez realizado este proceso con las tres enfermedades buscan los genes significativos comunes que existen en los tres. Con este proceso pretenden encontrar relación entre el trastorno de bipolaridad, la depresión y la esquizofrenia, ya que el primero presenta características comunes con los dos siguientes.

El objetivo de este trabajo es buscar si existen diferencias significativas entre los cuatro grupos, las diferentes enfermedades y los controles, a partir de los genes medidos y recogidos en dicha base de datos, utilizando el PERMANOVA y el BOOTMANOVA. Se pretende con esto encontrar genes que diferencien estas graves afecciones mentales.

Se tomaron muestras de 12625 genes en 50 cortezas prefrontales.

5.1.2. Resultados

Comenzamos haciendo un análisis previo de los datos para observar su distribución, como los datos no están normalizados realizaremos la normalización antes de empezar a trabajar con la matriz para eliminar la variabilidad adquirida al realizar el experimento y utilizar solo aquella propia de los datos. En la figura 5.1 se encuentran los boxplot de la base de datos sin normalizar (izquierda) y los de la base normalizada (derecha).

Se identificarán el número de individuos que pertenecen a cada grupo obteniendo 11 personas que padecían trastorno de bipolaridad, otros 11 con depresión, 13 con esquizofrenia y 15 controles.

El primer paso para realizar cualquiera de las técnicas descritas en el apartado teórico del trabajo (capítulo 3) es calcular las distancias. Una vez obtenida la matriz de distancias procederemos a realizar el PERMANOVA y el BOOTMANOVA.

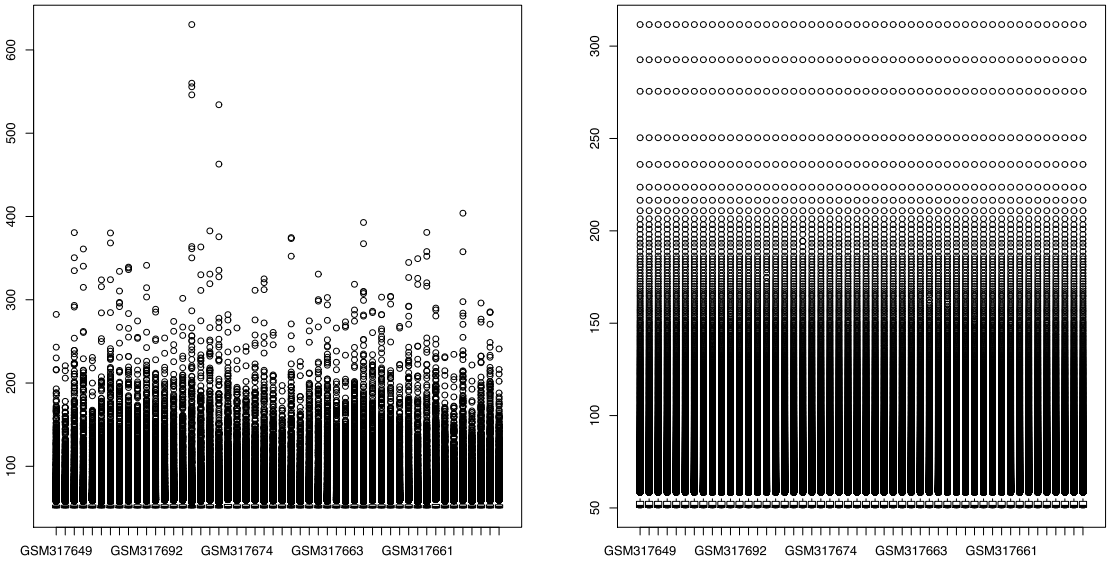

Figura 5.1

PERMANOVA

Al realizar el contraste general con 1000 permutaciones se obtiene que los resultados no son significativos, es decir no encuentra evidencias suficientes para afirmar que existen diferencias entre los valores de expresión génica de las diferentes enfermedades y de los controles. Los valores para el contraste general se pueden encontrar en el cuadro 5.1.

Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
269542.34	3919273	3	46	1.0545279	0.3756244

Cuadro 5.1: PERMANOVA General de las enfermedades mentales

Los contrastes para los controles y para cada una de las enfermedades, igual que en el contraste general, no se puede afirmar que existan diferencias.

Como describíamos en el apartado 5.1.1, en el artículo original realizan una comparación de cada uno de los grupos de enfermos con el grupo de controles para buscar genes que no son comunes a ambos. Por ello vamos a buscar si existen diferencias significativas utilizando un procedimiento similar al del artículo original.

Para llevar a cabo este proceso será necesario crear una matriz de contrastes que tenga la siguiente

forma

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \tag{5.1}$$

si introducimos esta matriz en el modelo obtenemos que, como era de esperar, ninguno de los resultados es significativo, se reflejan los valores del modelo en el cuadro 5.2.

Comparación	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
Bipolaridad	97374.92	3919273	1	46	1.1428770	0.2845715
Depresión	59478.67	3919273	1	46	0.6980935	0.7487251
Esquizofrenia	112688.74	3919273	1	46	1.0545279	0.1783822

Cuadro 5.2: PERMANOVA con las comparaciones de cada enfermedad con los controles

Podemos concluir que los resultados del estudio de Iwamoto et al. (2004) no pueden ser corroborados utilizando el PERMANOVA.

BOOTMANOVA

Para el BOOTMANOVA cogeremos 1000 muestras de los individuos para buscar diferencias entre los grupos. El resultado como se puede observar en el cuadro 5.3 es análogo al obtenido en el PERMANOVA.

Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
269542.34	3919273	3	46	1.0545279	0.3856144

Cuadro 5.3: PERMANOVA General de las enfermedades mentales

Volvemos a encontrar que no existen evidencias suficientes para demostrar que se observan diferencias significativas entre los grupos.

Igual que en el caso anterior realizaremos las comparaciones de cada una de las enfermedades con los controles para buscar diferencias entre ellos. La matriz de contrastes utilizada volverá a ser la (5.1). Los valores obtenidos se encuentran en el cuadro 5.4

La conclusión también es análoga a la anterior, no se pueden demostrar estadísticamente los resultados del artículo original utilizando el BOOTMANOVA.

Comparación	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
Bipolaridad	97374.92	3919273	1	46	1.1428770	0.2842716
Depresión	59478.67	3919273	1	46	0.6980935	0.7470253
Esquizofrenia	112688.74	3919273	1	46	1.3226133	0.1843816

Cuadro 5.4: PERMANOVA con las comparaciones de cada enfermedad con los controles

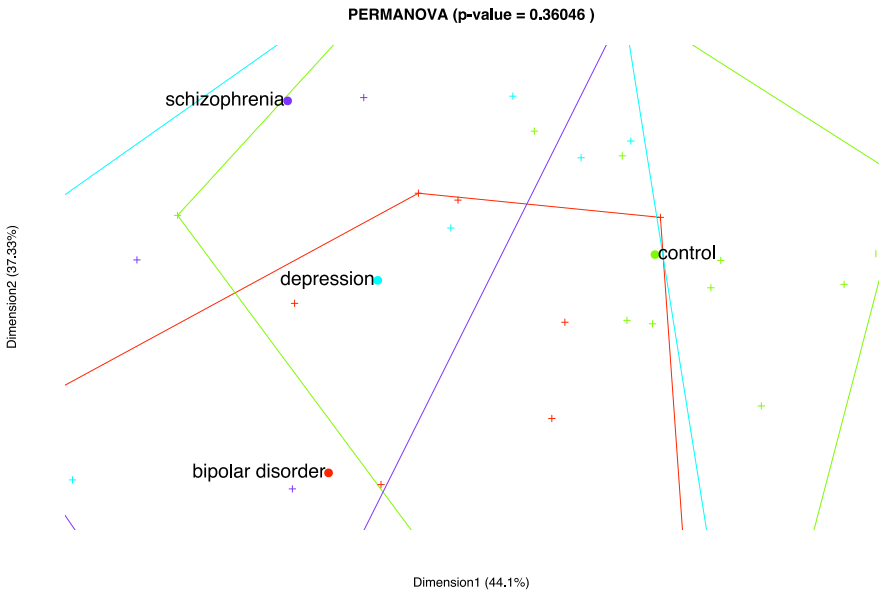

Figura 5.2: Análisis de Coordenadas Principales de las Enfermedades Mentales

Representación gráfica

Comenzaremos realizando una representación gráfica del Análisis de Coordenadas Principales para todos los individuos, calculada a partir de la matriz de distancia.

En esta representación del ACoP (Figura 5.2) podemos observar como todas las categorías quedan mezcladas y no se pueden observar ninguna de ellas diferenciada. Es decir, se observa gráficamente lo que habíamos concluido teóricamente con cualquiera de las dos técnicas.

A continuación vamos a crear una regiones de confianza a partir del punto central de cada uno de los grupos, para ello utilizaremos técnicas bootstrap como ha sido descrito en el apartado 4.3. En la figura 5.3 encontramos la representación con las tres primeras dimensiones calculadas en el Análisis Canónico Bootstrap realizado sobre la matriz de distancias.

Igual que en el caso anterior observamos que todos los grupos están muy próximos, es decir,

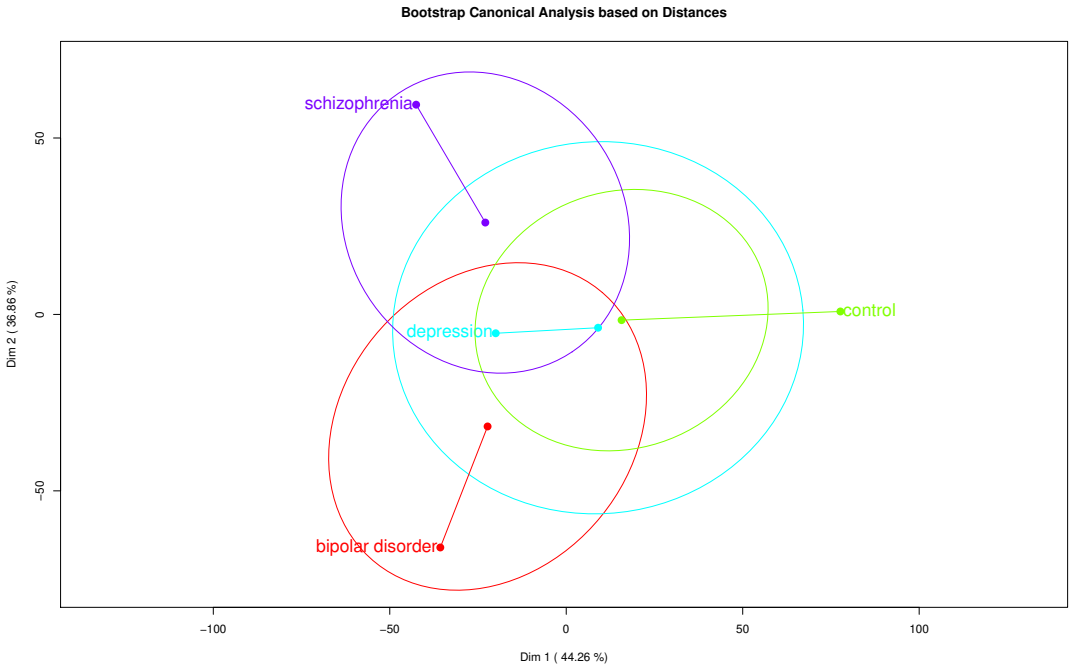

Figura 5.3: El rojo corresponde con el desorden de bipolaridad, el verde con los controles, azul depresión y morado esquizofrenia

corresponde con los resultados obtenidos de forma analítica.

5.2. Envejecimiento de la región cortical frontal del cerebro

5.2.1. Descripción de los datos

Existen diversos estudios (Chong et al., 1995; Ballas et al., 2005) que afirman que el represor neuronal REST está implicado en el desarrollo embrionario y, una vez que se regula, se encarga de la diferenciación neuronal terminal. El artículo del que hemos tomado los datos (Lu et al., 2014) tiene como objetivo mostrar que este represor neuronal también está implicado en el envejecimiento del cerebro humano y en la regulación de un conjunto de genes que influyen en la muerte celular, la resistencia al estrés y en la Enfermedad de Alzhéimer.

Este conjunto de muestras está recogido en la serie con clave de identificación GSE53890. Contiene 54675 perfiles de expresión génica relacionados con el represor neuronal REST, medidos en 41 regiones corticales frontales de individuos adultos:

- 12 jóvenes, por debajo de los 40 años

- 9 individuos de mediana edad, entre 40 y 70 años
- 16 ancianos normales, entre 70 y 94 años
- 4 extremadamente ancianos que están entre los 95 y los 106 años.

El objetivo de nuestro trabajo consistirá en buscar diferencias significativas entre los grupos de edad descritos anteriormente para comprobar si es cierto las afirmaciones de los estudios anteriores, es decir, que el represor neuronal REST se expresa de forma diferente dependiendo de la edad del sujeto.

5.2.2. Resultados

Obtendremos de la plataforma GEO los datos de ADN correspondientes a los 41 individuos. Les aplicaremos la transformación necesaria para obtener la expresión génica que emplearemos para nuestros análisis una vez que sean normalizados.

En la figura 5.4 encontramos los boxplot de los datos normalizados a la derecha y antes de su normalización a la izquierda.

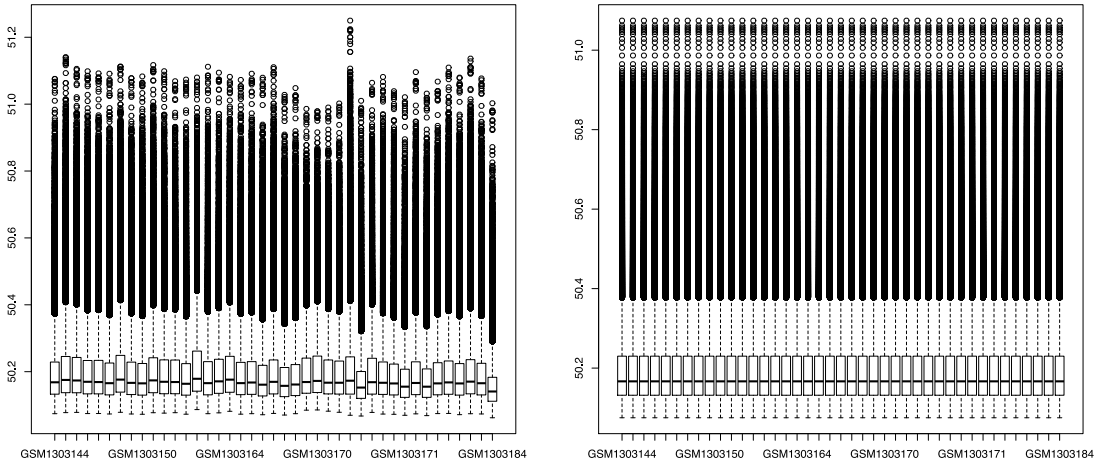

Figura 5.4

Tras el cálculo de la matriz de distancias e introducir el vector que determina los grupos, procedemos a realizar los cálculos del PERMANOVA y el BOOTMANOVA.

PERMANOVA

Comenzaremos realizando el contraste general para estudiar si existen diferencias significativas entre los grupos de edad, este contraste se encuentra recogido en el cuadro 5.5. Además, en esta misma tabla se pueden observar los contrastes individuales de cada uno de los grupos, contrastaremos si su media es igual 0.

Contraste	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
< 40 años	71.24228	685.8744	1	37	3.843217	0.000999001
40 – 70 años	20.52347	685.8744	1	37	1.107154	0.152847153
70 – 94	71.59585	685.8744	1	37	3.862291	0.000999001
94 – 106	41.35612	685.8744	1	37	2.230987	0.030969031
TOTAL	204.71773	685.8744	3	37	3.681216	0.000999001

Cuadro 5.5: PERMANOVA general del envejecimiento de la región cortical frontal del cerebro

En este caso sí que se encuentran diferencias altamente significativas entre los grupos, si realizamos los contrastes Post Hoc, no se pueden encontrar diferencias entre los jóvenes menores de 40 años y los adultos de mediana edad, sin embargo sí que encontramos diferencias estadísticamente significativas entre los grupos de mediana edad y los ancianos normales y entre estos últimos y los extremadamente ancianos, y diferencias altamente significativos entre todas las demás parejas de grupos.

Los grupos los podemos realizar incluyendo también el sexo e introducir en el modelo los dos efectos y la interacción entre ellos. Para ello debemos construir una matriz C de la siguiente forma

$$\mathbf{C} = \begin{pmatrix} 3 & 3 & -1 & -1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 2 & 2 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 3 & -3 & -1 & 1 & -1 & 1 & -1 & 1 \\ 0 & 0 & 2 & -2 & -1 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 \end{pmatrix} \tag{5.2}$$

Los tres primeros contrastes son los correspondientes a los grupos de edad, el siguiente pertenece al sexo y los tres últimos corresponden con la interacción.

Comenzaremos estudiando los contrastes de los efectos y el contraste general, estos resultados se encuentran en el cuadro 5.6

Efecto	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
Edad	197.38169	580.9724	3	33	3.737180	0.00009999
Sexo	38.35928	580.9724	1	33	2.178858	0.02849715
Interación	73.87875	580.9724	3	33	1.398804	0.08999100
TOTAL	309.61973	580.9724	7	33	2.512401	0.00009999

Cuadro 5.6: PERMANOVA de los efectos para el envejecimiento de la región cortical frontal del cerebro

Observamos que la interacción no es significativa, sin embargo en los efectos principales se encuentran diferencias estadística y altamente significativas para el sexo y la edad respectivamente.
Es posible estudiar los contrastes para cada uno de los efectos principales que han salido significativos, los resultados se encuentran en el cuadro 5.7

Contraste	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
Edad 1	87.84420	580.9724	1	33	4.9896667	0.00009999
Edad 2	64.29235	580.9724	1	33	3.6518908	0.00209979
Edad 3	45.24514	580.9724	1	33	2.5699838	0.01659834
Sexo	38.35928	580.9724	1	33	2.1788580	0.02849715

Cuadro 5.7: PERMANOVA de los efectos para el envejecimiento de la región cortical frontal del cerebro

En dicha tabla (cuadro 5.7) observamos que todos los contrastes son significativos, los dos primeros altamente significativos y los otros dos estadísticamente significativos, es decir, existen diferencias entre todos los grupos de edad y los valores de las variables varían en función del género de los individuos.

BOOTMANOVA

Podemos repetir el proceso realizado en el PERMANOVA, pero utilizando el BOOTMANOVA como técnica de comparación.

El contraste general se asemeja en gran medida al reflejado en el cuadro 5.5, los valores obtenidos para este caso se encuentran en el cuadro 5.8. Observaremos también cada uno de los grupos individuales para contrastar si la media de alguno de ellos es igual a 0.

Contraste	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
<40 años	71.24228	685.8744	1	37	3.843217	0.002997003
40 – 70 años	20.52347	685.8744	1	37	1.107154	0.160839161
70 – 94 años	71.59585	685.8744	1	37	3.862291	0.000999001
95 – 106 años	41.35612	685.8744	1	37	2.230987	0.023976024
TOTAL	204.71773	685.8744	3	37	3.681216	0.000999001

Cuadro 5.8: BOOTMANOVA general del envejecimiento de la región cortical frontal del cerebro

Algunos de los artículos que tratan sobre el represor neuronal REST afirman que una vez terminado el crecimiento dicho represor se estabiliza tomando valores más bajos y vuelve a expresarse en las personas de mayor edad interviniendo en el envejecimiento y la muerte neuronal, los contrastes individuales reflejados en el cuadro 5.8 refuerzan dicha afirmación ya que no existen evidencias suficientes para afirmar que los adultos de mediana edad (entre 40 y 70 años) tengan una expresión génica de la red de genes asociados al represor neuronal REST diferente de 0, mientras que en el resto de intervalos de edad sí que es posible encontrar diferencias significativas con el 0.

Incluiremos el sexo y realizaremos el estudio de los contrastes para examinar si, igual que en el caso anterior, son significativos.

Efecto	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
Edad	192.4618	580.9724	3	33	3.6440290	0.00019998
Sexo	27.9621	580.9724	1	33	1.5882842	0.08309169
Interación	29.1321	580.9724	3	33	0.5515806	0.70352965
TOTAL	309.61973	580.9724	7	33	2.512401	0.00039996

Cuadro 5.9: PERMANOVA de los efectos para el envejecimiento de la región cortical frontal del cerebro

En este caso obtenemos que solo la edad es significativa, es decir, solo se encuentran diferencias entre los diferentes grupos de edad. Los contrastes para estos grupos están recogidos en el cuadro 5.10.

Contraste	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
Edad 1	86.435856	580.9724	1	33	4.9096710	0.00059994
Edad 2	62.959276	580.9724	1	33	3.5761702	0.00419958
Edad 3	43.066704	580.9724	1	33	2.4462458	0.02279772

Cuadro 5.10: PERMANOVA de los efectos para el envejecimiento de la región cortical frontal del cerebro

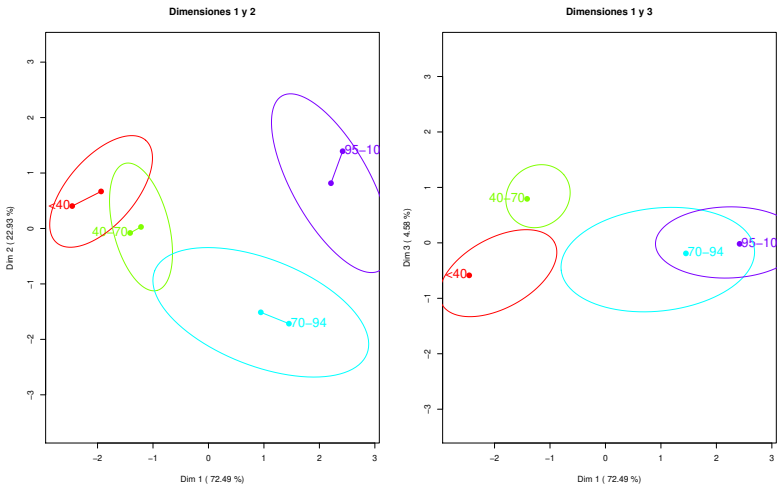

Figura 5.5: Análisis Canónico Bootstrap con los grupos de edad de envejecimiento de la región cortical frontal del cerebro

Todos los contrastes son significativos, es decir existen diferencias entre todos los grupos de edad. Al ser el único efecto significativo no será necesario estudiar el resto de los contrastes.

Representación gráfica

En este caso vamos a calcular el Análisis Canónico Bootstrap para la matriz de distancias con los grupos de edad (gráfico **??**).

Es posible observar la proximidad entre edades continúas, que algunas de ellas no presentaban significación, y a la vez ver las diferencias que existen entre los grupos de edad discontinuos.

Los gráficos creados a partir del ACoP utilizando como variables de agrupación la edad y el sexo no nos permiten encontrar todas las diferencias que se observaban analíticamente (figura 5.6).

Sin embargo los gráficos creados a partir del Análisis Canónico utilizando el bootstrap sí

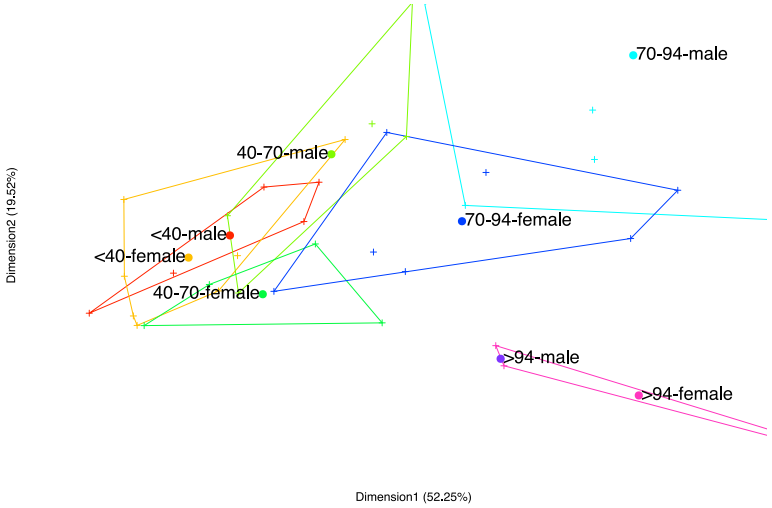

Figura 5.6: Análisis de Coordenadas Principales en el envejecimiento de la región cortical frontal del cerebro

que permiten encontrar las diferencias marcadas (figura 5.7). Los grupos de edades contiguas se encuentran más cercanos que los disjuntos, esto permite que los análisis localicen diferencias estadísticamente significativas entre los diferentes grupos de edad. Las diferencias que se observan de forma analítica dependiendo del género de los individuos muestreados se observan si examinamos el resto de planos calculados.

5.3. Enfermedad de Alzhéimer

5.3.1. Descripción de los datos

La tercera base de datos con la que vamos a trabajar pertenece a un estudio realizado por Konietschke et al. (2015) que trabaja sobre la Enfermedad de Alzhéimer, teniendo en cuenta también la Diabetes Mellitus como factor de riesgo.

La Enfermedad de Alzhéimer (EA) es una enfermedad neurodegenerativa que presenta un deterioro cognitivo y una serie de trastornos en las conductas. Una de las características más habituales de esta enfermedad es la pérdida de memoria inmediata. Hay estudios que afirman que la EA afecta a entre 25 y 44 millones de personas en el mundo.
Es una enfermedad bastante conocida para la que se sigue investigando un tratamiento efectivo que la cure, sin embargo hasta el momento, aunque los estudios realizados son muchos, no se ha

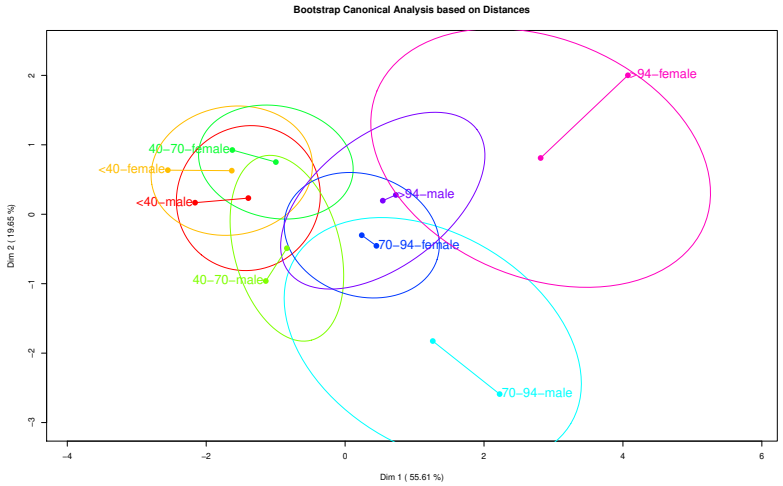

Figura 5.7: Análisis Canónico Bootstrap del envejecimiento de la región cortical frontal del cerebro

encontrado cura.

La Diabetes Mellitus es una enfermedad provocada por la insuficiencia de insulina producida por el páncreas, o porque esta no es eficiente en el organismo, que genera un exceso de glucosa en sangre. Esta enfermedad es considerada en diversos estudios como factor de riesgo de otras enfermedades o trastornos. Existen diferentes tipos de Diabetes Mellitus en función de la forma o el momento en el que el organismo deja de tener el autoabastecimiento necesario de insulina.

El conjunto de muestras relacionadas en este caso, también se han obtenido de GEO. La clave de localización de la serie es GSE36980.

Esta base de datos contiene la información de 33297 genes medidos en 79 individuos. Para extraer los datos se obtuvieron 88 muestras de ADN de la materia gris del cerebros postmorten de tres zonas diferentes, cortezas frontales, cortezas temporales e hipocampos. Tras realizar los exámenes de control de calidad pertinentes se han conservado 79 muestras, en el cuadro 5.11 se encuentran distribuidas en función de la zona cerebral a la que pertenecen y si tienen o no la EA o una enfermedad de características similares.

El artículo original utiliza estos datos para realizar un ANOVA de tres vías en el que uno de los factores es el lugar del cerebro donde se toma la muestra, otro si padecía o no la EA y demencia vascular y por último el sexo del individuo del que se tomó la muestra. En dicho artículo,

	Corteza frontal	Corteza temporal	Hipocampos	**TOTAL**
EA	15	10	7	32
no EA	18	19	10	47
TOTAL	33	29	17	**79**

Cuadro 5.11: Descripción de los datos de EA

también se contrasta con muestras tomadas de ratones transgénicos con EA con las citada anteriormente, se buscan genes relacionados con uno de los tipos de DM y la obesidad y diversos trastornos psiquiátricos con EA.

En nuestro caso vamos a utilizar los datos del cuadro 5.11, incluyendo como tercer factor el sexo, para estudiar si existen diferencias significativas entre todos los grupos. Emplearemos este ejemplo para ilustrar uno de los diseños de tres vías con interacción de los MANOVAs basados en distancias, desarrollado en el apartado teórico 3.4.3 para el caso del BOOTMANOVA, en el caso del PERMANOVA se realizará de forma análoga a la citada anteriormente.

5.3.2. Resultados

En este tercer ejemplo vamos a realizar un procedimiento similar a los dos casos anteriores (apartados 5.1 y 5.2). Extraeremos los datos de la plataforma GEO y realizaremos un boxplot de la expresión de los genes antes y después de su normalización (Figura 5.8)

En este caso tendremos tres factores de variación, si el individuo presenta o no la enfermedad, de qué parte del cerebro ha sido tomada la muestra y su género. Se creará un factor que identifique a los individuos de la muestra en función de dichas características.

A continuación, se detallarán los resultados del PERMANOVA y el BOOTMANOVA calculados a partir de la matriz de distancias observadas.

PERMANOVA

Igual que en el caso anterior (sección 5.2.2) vamos a estudiar el contraste general y para cada uno de los efectos. Será necesario construir una matriz C que contenga los contrastes para los tres factores y sus interacciones, tanto dobles como triples. La ecuación 5.3 contiene la matriz de

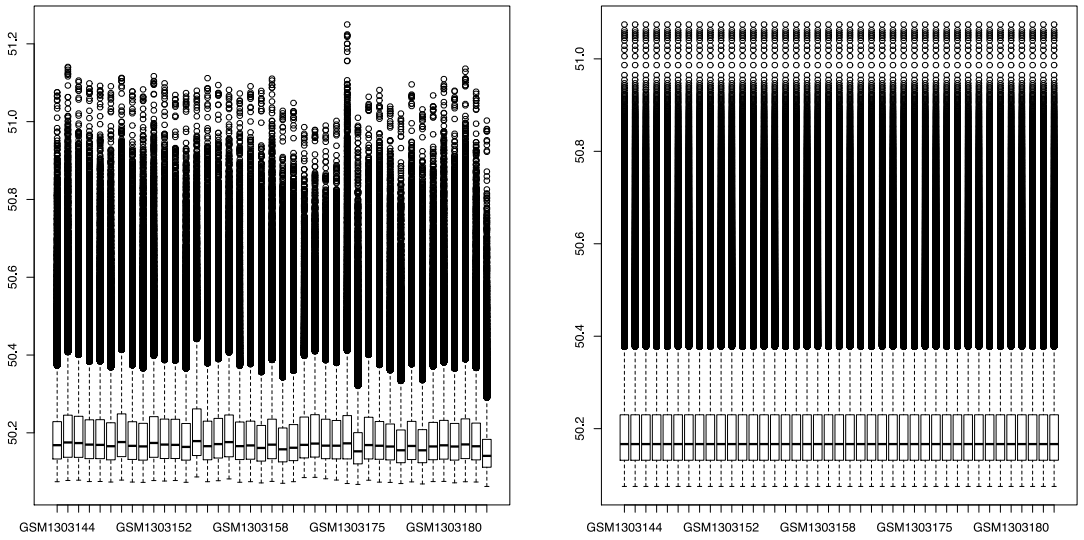

Figura 5.8

contrastes que utilizaremos construida como se explicó en la sección 2.5.

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 2 & 2 & -1 & -1 & -1 & -1 & 2 & 2 & -1 & -1 & -1 & -1 \\ 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 2 & 2 & -1 & -1 & -1 & -1 & -2 & -2 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 2 & -2 & -1 & 1 & -1 & 1 & 2 & -2 & -1 & 1 & -1 & 1 \\ 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 1 & -1 & -1 & 1 \\ 2 & -2 & -1 & 1 & -1 & 1 & -2 & 2 & 1 & -1 & 1 & -1 \\ 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & -1 & 1 & 1 & -1 \end{pmatrix} \tag{5.3}$$

Se formará el modelo para estudiar los tres efectos y sus interacciones, los resultados obtenidos están en el cuadro 5.12.

Se puede observar que los tres efectos son altamente significativos, mientras no existen evidencias suficientes para afirmar que alguna de las interacciones sea significativa. Realizaremos por tanto los contrastes para los tres efectos principales que serán recogidos en el cuadro 5.13.

Efecto	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
Enfermedad	8.189370e-07	2.574725e-05	1	67	2.1310539	0.00069993
Tejido	5.669296e-06	2.574725e-05	2	67	7.3763761	0.00009999
Sexo	1.061346e-06	2.574725e-05	1	67	2.7618551	0.00009999
Enfermedad*Tejido	7.119863e-07	2.574725e-05	2	67	0.9263723	0.66773323
Enfermedad*Sexo	4.831185e-07	2.574725e-05	1	67	1.2571804	0.10448955
Tejido*Sexo	7.183630e-07	2.574725e-05	2	67	0.9346691	0.64243576
Enfermedad*Tejido*Sexo	6.630314e-07	2.574725e-05	2	67	0.8626765	0.85781422
TOTAL	1.012608e-05	2.574725e-05	11	67	2.3954798	0.00009999

Cuadro 5.12: PERMANOVA de los efectos para la EA en función del tipo de tejido cerebral extraído y del sexo

Contraste	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
Enfermedad	8.189370e-07	2.574725e-05	1	67	2.1310539	0.00069993
Tejido 1	2.438978e-06	2.574725e-05	1	67	6.3467548	0.00009999
Tejido 2	3.230318e-06	2.574725e-05	1	67	8.4059973	0.00009999
Sexo	1.061346e-06	2.574725e-05	1	67	2.7618551	0.00009999

Cuadro 5.13: PERMANOVA de los efectos para el envejecimiento de la región cortical frontal del cerebro

Los contrastes permiten afirmar que existen diferencias significativas entre todos los grupos de los tres factores principales, es decir, existen diferencias en la expresión génica dependiendo de si el individuo padece EA o no, también se encuentran valores diferentes dependiendo de la zona del cerebro de donde ha sido extraído el tejido. Además, dependiendo de si el individuo es hombre o mujer, también es posible observar diferencias.

BOOTMANOVA

De forma análoga a los casos anteriores, repetiremos el proceso empleando el BOOTMANOVA, comenzaremos recogiendo en el cuadro 5.14 los valores para el contraste general y el estudio de los tres efectos y su correspondiente interacción.

Obtenemos el mismo resultado que en el caso del PERMANOVA de esta sección (5.3.2) , únicamente los efectos principales son significativos. Realizaremos, por tanto, los contrastes para dichos efectos (Cuadro 5.15).

Efecto	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
Enfermedad	8.189370e-07	2.574725e-05	1	67	2.1310539	0.00279972
Tejido	5.669296e-06	2.574725e-05	2	67	7.3763761	0.00009999
Sexo	1.061346e-06	2.574725e-05	1	67	2.7618551	0.00029997
Enfermedad*Tejido	7.119863e-07	2.574725e-05	2	67	0.9263723	0.61363864
Enfermedad*Sexo	4.831185e-07	2.574725e-05	1	67	1.2571804	0.14788521
Tejido*Sexo	7.183630e-07	2.574725e-05	2	67	0.9346691	0.60533947
Enfermedad*Tejido*Sexo	6.630314e-07	2.574725e-05	2	67	0.8626765	0.76522348
TOTAL	1.012608e-05	2.574725e-05	11	67	2.3954798	0.00009999

Cuadro 5.14: BOOTMANOVA de los efectos para la enfermedad de Alzhéimer

Contraste	Explicada	Residual	G.L. Num	G.L. Denom	F-exp	p-valor
Enfermedad	8.189370e-07	2.574725e-05	1	67	2.1310539	0.00279972
Tejido 1	2.438978e-06	2.574725e-05	1	67	6.3467548	0.00009999
Tejido 2	3.230318e-06	2.574725e-05	1	67	8.4059973	0.00009999
Sexo	1.061346e-06	2.574725e-05	1	67	2.7618551	0.00029997

Cuadro 5.15: PERMANOVA de los efectos para el envejecimiento de la región cortical frontal del cerebro

Podemos concluir que todos los contrastes sometidos a estudio son significativos, por lo tanto, existen diferencias significativas en cada uno de los efectos por separado.

Representación gráfica

Para este ejemplo ambas representaciones diferencian con gran claridad los grupos que analíticamente se habían descrito como diferentes. En la figura 5.9 observaremos los cuatro primero planos factoriales en los que se pueden observar la mayor parte de los grupos diferenciados. En el primer plano factorial la separación en función del tejido del que se ha extraído la muestra está claramente diferenciada.

En el caso del Análisis Canónico Bootstrap de esta matriz de distancias los resultados de la representación son muy similares. Utilizaremos estrellas para presentar a cada uno de los grupos uniendo cada punto con la media para la representación del plano formado por las dos primeras dimensiones. Este gráfico se puede observar en la figura 5.10. En él se ven tres agrupaciones claramente diferenciadas, cada una de ellas corresponde con un tipo de tejido del que se ha obtenido la muestra, hipocampo, corteza temporal y corteza frontal.

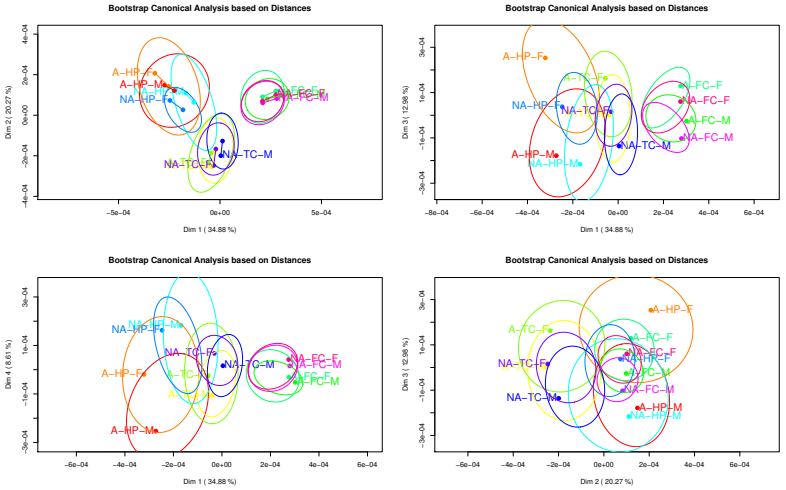

Figura 5.9: Los cuatro primeros planos del Análisis Canónico Bootstrap del estudio sobre la Enfermedad de Alzhéimer

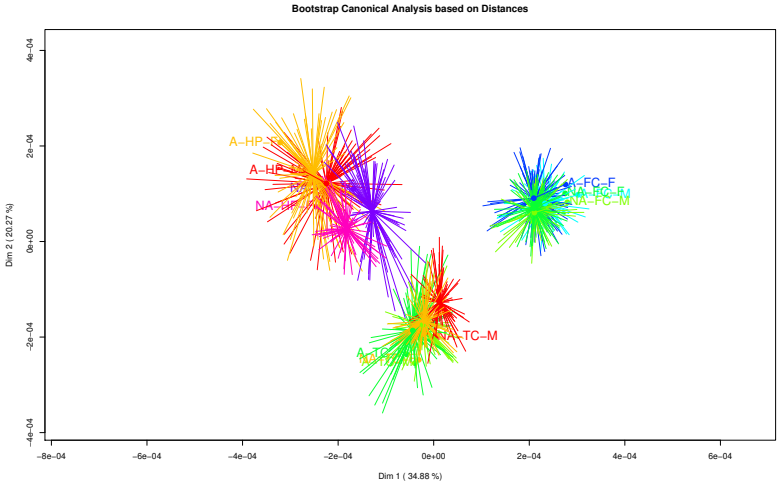

Figura 5.10: Análisis Canónico Bootstrap del estudio sobre la Enfermedad de Alzhéimer

5.4. HapMap

5.4.1. Descripción de los datos

El último conjunto de datos que vamos a utilizar no está asociado a ninguna enfermedad, aunque sí a la genética de los seres vivos. En este caso no se cuantificará la intensidad lumínica proyectada por el microarray, sino que únicamente se observará si está presente o ausente los polimorfismos mononucleotídico del gen en la propia cadena de ADN.

Vamos a utilizar esta base de datos para buscar diferencias significativas entre poblaciones. Dentro del ADN todos los seres humanos compartimos el 99,9 % de nuestra información genética y los cambios genotípicos observados entre los individuos están dentro del 0,1 % restante, el color del pelo, los ojos o la piel son algunos de ellos, aunque también pueden incluirse riesgos para padecer una enfermedad o el grupo sanguíneo.
Es posible encontrar secuencias de nucleótidos dentro del ADN donde cambia una única base, en estos casos se trata de polimorfismos mononucleotídicos o SNP ("Single Nucleotide Polymorphism"). Denominamos haplotipo a un gran número de SNPs que siguen un patrón y constituyen un único bloque. La recombinación genética dentro de un haplotipo es muy baja, por lo tanto, generalmente se hereda la SNPs completa.

El proyecto HapMap comenzó con la intención de crear un mapa de haplotipos del genoma humano, para ello en 2002 se realizó una reunión en la que se decidió muestrear una serie de individuos, este muestreo fue llevado a cabo un centro de investigación en el que estaban incluidos cinco países (Reino Unido, Canadá, Japón, China, Nigeria y Estados Unidos).
Dicho proyecto fue dividido en tres etapas en cada una de ellas se amplían los resultados obtenidos.

Fase I: En esta primer etapa se encuentran más de un millón de resultados que fueron publicados en el año 2005. Para ello se recoge una muestra de 269 personas de Nigeria, Japón, China y Estados Unidos (diferenciando el linaje de ascendencia) en representación de la población mundial.

Fase II: En 2007 se publican los nuevos resultados obtenidos en esta segunda etapa. La muestra

continúa siendo la misma, pero se pretenden encontrar un mayor número de SNP (3,2 millones) (International HapMap Consortium, 2007; Skipper, 2007).

Fase III: En la última fase del proyecto se aumentan las poblaciones de las 5 de la fase I a 11. Los SNP añadidos como resultados al estudio fueron 1,6 millones y fueron publicados en el año en 2009 .

Los SNPs que identifican al haplotipo se denominan tag SNPs y corresponden con los resultados que se han presentado en el proyecto. En total, los resultados originales, obtienen en torno a 10 millones de SNPs de los cuales 500000 son tags SNPs.

Los datos que vamos a utilizar corresponden con las 11 poblaciones de la Fase III. Dichas áreas poblacionales se han recogido en el cuadro 5.16 donde se han obtenido las muestras.

Código	**Población**
CEU	Utah con ascendencia Europa del norte y occidental.
CHB	Chinos Han de Beijing, China.
JPT	Japoneses de Tokyo, Japón.
YRI	Yoruba de Ibadán, Nigeria.
ASW	Estadounidenses del suroeste con ascendencia africana.
CHD	Chinos en la metrópolis de Denver, Colorado, Estados Unidos.
GIH	Indios Gujarati residentes en Houston, Texas, Estados Unidos.
LWK	De étnia Luhya de Webuye, Kenia.
MKK	Masáis de Kinyawa, Kenia.
MEX	De Los Ángeles, California, Estados Unidos con ascendencia mejicana.
TSI	Residentes en la Toscana de Italia.

Cuadro 5.16: Poblaciones de la fase III del proyecto HapMap con la codificación realizada en la aplicación práctica.

Todos los datos recogidos llevan asociado un compromiso y consentimiento informado internacional del proyecto (Rotimi et al., 2007) así como un estudio ético de la investigación (International HapMap Consortium and others, 2004).

Más información sobre este proyecto se puede obtener en los artículos asociados al HapMap de la bibliografía (International HapMap Consortium and others, 2005; McVean et al., 2005). También existen algunos artículos que presentan su disconformidad con dicho proyecto (Terwilliger and Hiekkalinna, 2006).

La utilización de los datos generados por este estudio ha sido muy diversa (Manolio et al., 2008; Bell et al., 2011; McVean et al., 2005; Deloukas and Bentley, 2004; Gitschier, 2009; Thorgeirsson et al., 2008; Smyth et al., 2006), algunas de las aplicaciones más habituales es la búsqueda de genotipos o SNP asociados a enfermedades.

5.4.2. Resultados

Antes de realizar los análisis se han extraído únicamente los polimorfismos del cromosoma 10 y que no tienen datos perdidos resultando una matriz de 1397 individuos en los que se han observado 8384 alelos correspondientes a 4192 polimorfismos.
La medida de similaridad utilizada para calcular las matriz de distancias es el coeficiente de concordancia de Simple, $\frac{a+d}{a+b+c+d}$.

PERMANOVA

Realizaremos el contraste general para buscar diferencias significativas entre las poblaciones descritas en el cuadro 5.16. Los resultados del contraste se pueden encontrar en el cuadro 5.17. Observamos que son altamente significativas las diferencias entre los grupos, además en el cuadro 5.18 podemos observar que los contrastes individuales también son altamente significativos.

	Explained	Residual	G.L. Num	G.L. Denom	F-exp	p-value
Total	28.41	183.78	10	1386	21.43	0.00

Cuadro 5.17: PerMANOVA

BOOTMANOVA

Para el caso del BOOTMANOVA repetiremos el mismo contraste general (5.19) que en el caso anterior para intentar identificar si existen diferencias significativas entre los diferentes grupos étnicos muestreados.

Los resultados obtenidos nos permiten comprobar que sí que existen diferencias entre los grupos y que son altamente significativas, además realizaremos los contrastes individuales para cada uno de los grupos. En la tabla 5.20 observamos que todos los contrastes individuales resultan

	Explained	Residual	G.L. Num	G.L. Denom	F-exp	p-value
C ASW	1.05	183.78	1	1386	7.92	0.00
C CEU	2.89	183.78	1	1386	21.78	0.00
C CHB	3.38	183.78	1	1386	25.52	0.00
C CHD	2.84	183.78	1	1386	21.45	0.00
C GIH	1.59	183.78	1	1386	11.96	0.00
C JPT	2.96	183.78	1	1386	22.30	0.00
C LWK	2.51	183.78	1	1386	18.94	0.00
C MEX	1.60	183.78	1	1386	12.10	0.00
C MKK	2.48	183.78	1	1386	18.71	0.00
C TSI	1.75	183.78	1	1386	13.22	0.00
C YRI	5.35	183.78	1	1386	40.36	0.00
Total	28.41	183.78	10	1386	21.43	0.00

Cuadro 5.18: Contrastes del PERMANOVA para el proyecto HapMap

	Explained	Residual	G.L. Num	G.L. Denom	F-exp	p-value
Total	28.41	183.78	10.00	1386.00	21.43	0.00

Cuadro 5.19: MANOVA Bootstrap basado en distancias para el proyecto HapMap

altamente significativos, es decir ninguno de los grupos se puede afirmar que sea igual a cero.

Representación gráfica

En este caso realizaremos la representación gráfica correspondiente con un Análisis Canónico Bootstrap.

Se puede observar que, al tener una gran muestra, los resultados gráficos son más claros que en los ejemplos anteriores. La figura 5.11 nos permite corroborar los resultados analíticos calculados en los apartados anteriores, observando los diferentes planos podemos diferenciar cada uno de los grupos con una confianza del 95 %.

En el plano formado por las dos primeras dimensiones, que recoge el 82,47 % de la información, se observan diferentes agrupaciones entre los individuos de las distintas poblaciones. La elipse de las medias verde, correspondiente con los individuos de Yoruba de Ibadán, Nigeria (YRI), y la

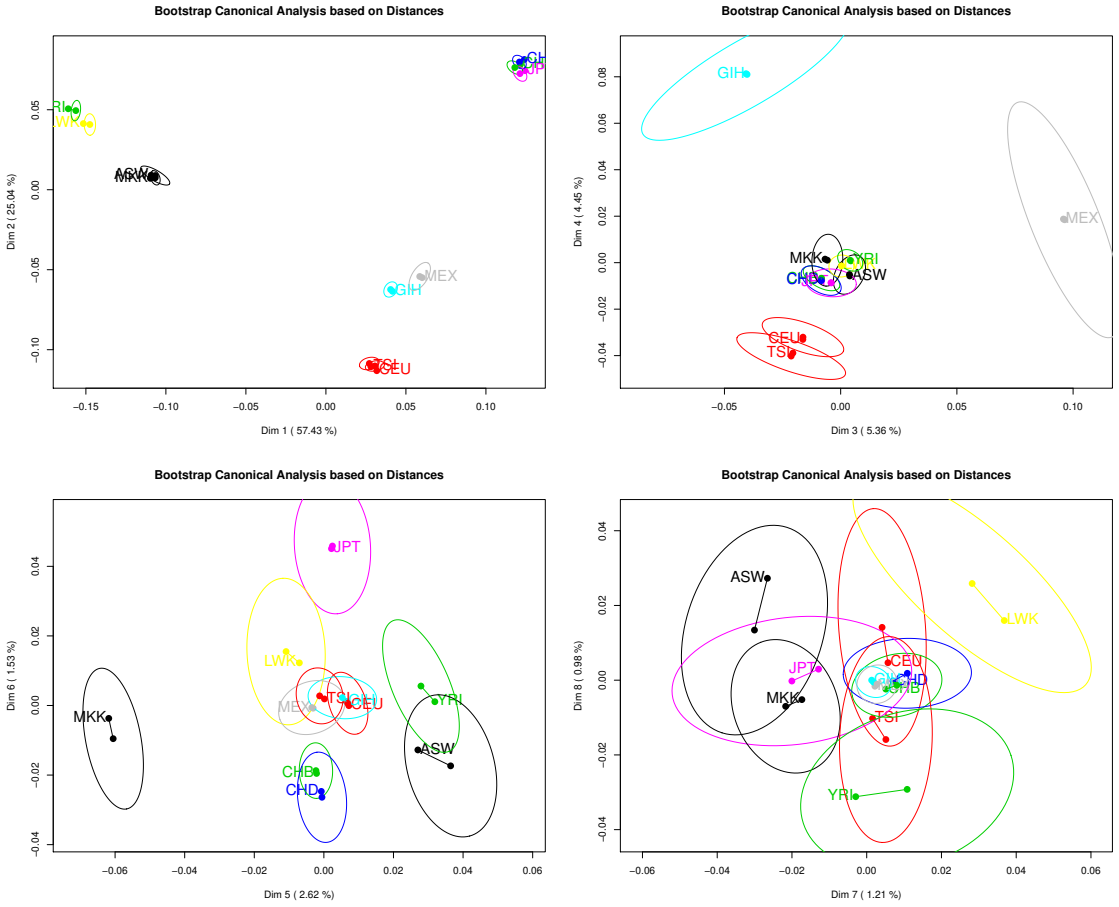

Figura 5.11: Análisis Canónico Bootstrap del Proyecto Hap Map

	Explained	Residual	G.L. Num	G.L. Denom	F-exp	p-value
C ASW	1.05	183.78	1.00	1386.00	7.92	0.00
C CEU	2.89	183.78	1.00	1386.00	21.78	0.00
C CHB	3.38	183.78	1.00	1386.00	25.52	0.00
C CHD	2.84	183.78	1.00	1386.00	21.45	0.00
C GIH	1.59	183.78	1.00	1386.00	11.96	0.00
C JPT	2.96	183.78	1.00	1386.00	22.30	0.00
C LWK	2.51	183.78	1.00	1386.00	18.94	0.00
C MEX	1.60	183.78	1.00	1386.00	12.10	0.00
C MKK	2.48	183.78	1.00	1386.00	18.71	0.00
C TSI	1.75	183.78	1.00	1386.00	13.22	0.00
C YRI	5.35	183.78	1.00	1386.00	40.36	0.00
Total	28.41	183.78	10.00	1386.00	21.43	0.00

Cuadro 5.20: Contrastes para el MANOVA Bootstrap basado en distancias del proyecto HapMap

amarilla, correspondientes a los individuos de etnia Luhya de Webuye, Kenia, podría definirse como un grupo que recibe los valores más bajos en la primera dimensión y altos en la segunda; no se encuentran muy distantes de ellos los dos grupos representados en color negro, que corresponden con los estadounidenses del suroeste con ascendencia africana (ASW) y los masáis de Kinyawa, Kenia (MKK), estos cuatro grupos tienen ascendencia africana. Los valores más altos en las dos primeras dimensiones corresponden con el grupo verde, chinos Han de Beijing, China (CHB), el azul, chinos en la metrópolis de Denver, Colorado, Estados Unidos (CHD) y el rosa, japoneses de Tokyo, Japón (JPT), todos ellos corresponden con los individuos de la muestra que tienen ascendencia asiática. En color rojo se presentan juntos, en cualquiera de los planos, otros dos grupos de individuos, los residentes en la Toscana de Italia (TRI) y los individuos de Utah con ascendencia Europa del norte y occidental (CEU), es decir, aquellas muestras procedentes de personas con ascendencia europea. Por último, en color gris encontramos individuos de Los Ángeles, California, Estados Unidos con ascendencia mejicana (MEX) y en azul claro los indios Gujarati residentes en Houston, Texas, Estados Unidos (GIH), en ambos casos corresponden con personas residentes en el sur de EE.UU. Esto nos permite comprobar que existen asociaciones entre las diferentes poblaciones con ascendencia similar, aunque sea posible diferenciar todos los grupos entre sí.

Capítulo 6

Conclusiones

- Se ha realizado el estudio del fundamento teórico del Análisis Multivariante de la Varianza (MANOVA) con base en el Modelo Lineal General Multivariante (MLG). Se detallan en el estudio las condiciones de aplicación de dicha técnica y la motivación para la realización de este trabajo debido a las numerosas restricciones que presenta.

- Se ha hecho una revisión de la metodología utilizada para extraer el efecto de cada uno de los factores o la interacción entre ellos empleando la matriz de contrastes **C**. De forma simultánea al estudio de esta matriz (**C**), se realiza el análisis de la matriz **M** que permite la comparación y la búsqueda de diferencias entre las variables o combinaciones de las mismas. Para el desarrollo de las formas no paramétricas será necesario el conocimiento detallado de la teoría clásica de los MLG.

- Se ha llevado a cabo una revisión de las técnicas no paramétricas que permiten el estudio de la diferenciación entre grupos en los que el número de variables respuesta es bastante mayor que el número de individuos.

- Se ha profundizado en una de las técnicas no paramétricas denominada PERMANOVA (Anderson, 2001; McArdle and Anderson, 2001) con buenos resultados de aplicación, pero que puede presentar algunos problemas en conjuntos de datos grandes.

- Se ha detallado una técnica fundamentada en principios similares al PERMANOVA, pero que solventa los problemas presentados por esta. Las dos técnicas tienen en común la realización de MANOVAs basados en distancias.

- Se han recogido varias medidas de distancia o disimilitud para los diferentes tipos de variables o combinación de ellas.

- Se ha desarrollado el fundamento teórico del BOOTMANOVA propuesto como alternativa al PERMANOVA detallando previamente el modelo inicial. El fundamento teórico permite incluir matrices de combinaciones entre individuos y/o entre variables para la construcción de modelos más complejos, análogos a los creados en el MANOVA, base de la técnica BOOTMANOVA.

- Se ha utilizado un Análisis de Coordenadas Principales sobre los centroides para generar una representación gráfica en baja dimensión de los resultados obtenidos tanto en el PERMANOVA como en el BOOTMANOVA, similar a la que proporciona el Análisis Canónico con los resultados del MANOVA.

- Se propone la representación gráfica de los resultados del BOOTMANOVA mediante Coordenadas Principales de la matriz de centroides y se calculan regiones de confianza empleando basadas en técnicas bootstrap. Hemos denominado a la técnica *Análisis Canónico Bootstrap* sobre la matriz de distancias.

- Se muestran cuatro casos prácticos de aplicación de las técnicas PERMANOVA y BOOTMANOVA con datos pertenecientes al genoma humano, y que presentan un mayor número de diferencias que los métodos utilizados hasta ahora.

- En el primer ejemplo práctico se concluye que no se puede afirmar que los resultados obtenidos por Iwamoto et al. (2004) sean ciertos utilizando los dos métodos de MANOVAs basados en distancias. El PERMANOVA y el BOOTMANOVA concluyen que no existen diferencias significativas entre las diferentes enfermedades y los controles, dicha conclusión se ve reflejada también en los gráficos asociados.

- El ejemplo práctico número dos nos permite concluir, en concordancia con varios artículos de referencia, que la expresión génica de los genes asociados al represor neuronal REST presentan diferencias en función de la edad y en los adultos de mediana edad (entre 40 y 70 años) no se encuentran evidencias para afirmar que sea diferente de 0. También es posible encontrar diferencias en función del sexo, empleando la nueva representación gráfica explicada en la sección 4.3 se pueden observar dichos resultados de forma gráfica con mayor claridad.

- En el ejemplo tres, relacionado con la enfermedad de Alzhéimer, podemos observar que existen diferencias significativas en función del tejido del cerebro del cual ha sido extraída la muestra, del género del individuo y de si padecían o no la enfermedad de Alzhéimer. En este caso las diferentes representaciones gráficas explicadas tienen un comportamiento muy similar.

- El último ejemplo se ha realizado con una matriz de datos binarios perteneciente al proyecto HapMap obteniendo mayores diferenciaciones significativas que en el artículo original con las dos técnicas aplicadas. Para el estudio de la hipótesis alternativa se ha realizado la representación gráfica del Análisis Canónico Bootstrap que permite encontrar asociaciones entre poblaciones con una ascendencia similar en el plano formado por las dos primeras dimensiones, los siguientes planos permiten diferenciar cada una de las poblaciones de todas las demás.

- Se ha extendido para variables binarias el método de Gower and Krzanowski (1999), en dichas variables no es posible el calculo del centroide de forma directa, lo que no permite calcular directamente la distancia de cada punto al centro. Para que sea posible realizar dichos cálculos se emplean las coordenadas principales de la matriz completa.

- Se han comparado los resultados del PERMANOVA y el BOOTMANOVA observando que son prácticamente idénticas, pero en los casos en los que existen número de individuos elevado, el BOOTMANOVA tiene una mayor potencia.

Bibliografía

Amaro, R., Vicente-Villardon, J. L., and Galindo Villardón, M. P. (2008). Contribuciones al manova-biplot: regiones de confianza alternativas. *Revista de Investigación Operacional*, 29:231–241.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1):32–46.

Anderson, M. J. (2005). Permanova: a fortran computer program for permutational multivariate analysis of variance. *Department of Statistics, University of Auckland, New Zealand*, 24.

Anderson, M. J. and Willis, T. J. (2003). Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. *Ecology*, 84(2):511–525.

Arthur, J. C., Perez-Chanona, E., Mühlbauer, M., Tomkovich, S., Uronis, J. M., Fan, T.-J., Campbell, B. J., Abujamel, T., Dogan, B., Rogers, A. B., Rhodes, J. M., Stintzi, A., Simpson, K. W., Hansen, J. J., Keku, T. O., Fodor, A. A., and Jobin, C. (2012). Intestinal Inflammation Targets Cancer-Inducing Activity of the Microbiota. *Science*, 338(6103):120–123.

Ballas, N., Grunseich, C., Lu, D. D., Speh, J. C., and Mandel, G. (2005). Rest and its corepressors mediate plasticity of neuronal gene chromatin throughout neurogenesis. *Cell*, 121(4):645–657.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., and Pritchard, J. K. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, 12:R10.

Chapman, M. G. and Underwood, A. J. (1999). Ecological patterns in multivariate assemblages: information and interpretation of negative values in ANOSIM tests. *Marine Ecology Progress Series*, 180:257–265.

Cheng, R. and Palmer, A. A. (2013). A simulation study of permutation, bootstrap, and gene dropping for assessing statistical significance in the case of unequal relatedness. *Genetics*, 193(3):1015–8.

Chong, J. A., Tapia-Ramirez, J., Kim, S., Toledo-Aral, J. J., Zheng, Y., Boutros, M. C., Altshuller, Y. M., Frohman, M. A., Kraner, S. D., and Mandel, G. (1995). Rest: a mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell*, 80(6):949–957.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, 18(1):117–143.

Cuadras Avellanas, C. (1988). Distancias estadísticas. *Estadística Española*, (119):295–357.

De Leeuw, J. and Meulman, J. (1986). A special jackknife for multidimensional scaling. *Journal of Classification*, 3(1):97–112.

Deloukas, P. and Bentley, D. (2004). The HapMap project and its application to genetic studies of drug response. *The Pharmacogenomics Journal*, 4(2):88–90.

Duarte, L. C., Von Zuben, F. J. A., and Reis, S. A. F. d. (1998). Orthogonal projections and bootstrap resampling procedures in the study of intraspecific variation. *Genetics and Molecular Biology*, 21.

Efron, B. (1979). Bootstrap methods: another look at the jackknife annals of statistics 7: 1–26. *View Article PubMed/NCBI Google Scholar*.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Flather Curtis H. and Sauer John R. (1996). Using Landscape Ecology to Test Hypotheses About Large-Scale Abundance Patterns in Migratory Birds. *Ecology*, 77(1):28–35.

Gitschier, J. (2009). Inferential Genotyping of Y Chromosomes in Latter-Day Saints Founders and Comparison to Utah Samples in the HapMap Project. *The American Journal of Human Genetics*, 84(2):251–258.

Goodnight, C. J. and Schwartz, J. M. (1997). A bootstrap comparison of genetic covariance matrices. *Biometrics*, pages 1026–1039.

Goshtasby, A. A. (2012). Similarity and Dissimilarity Measures. In *Image Registration*, pages 7–66. Springer London, London.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Gower, J. C. (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika*, 55(3):582–585.

Gower, J. C. and Krzanowski, W. J. (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):505–519.

Gray, A. and Markel, J. (1976). Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):380–391.

Han, D., Dezert, J., Han, C., and Yang, Y. (2011). New Dissimilarity Measures in Evidence Theory. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–7. IEEE.

International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861.

International HapMap Consortium and others (2004). Integrating ethics and science in the International HapMap Project. *Nature reviews. Genetics*, 5(6):467–475.

International HapMap Consortium and others (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

Iwamoto, K., Kakiuchi, C., Bundo, M., Ikeda, K., and Kato, T. (2004). Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. *Molecular Psychiatry*, 9:406–416.

Konietschke, F., Bathke, A. C., Harrar, S. W., and Pauly, M. (2015). Parametric and nonparametric bootstrap methods for general MANOVA. *Journal of Multivariate Analysis*, 140:291–301.

Krishnamoorthy, K. and Lu, F. (2010). A parametric bootstrap solution to the manova under heteroscedasticity. *Journal of Statistical Computation and Simulation*, 80(8):873–887.

Lu, T., Aron, L., Zullo, J., Pan, Y., Kim, H., Chen, Y., Yang, T.-H., Kim, H.-M., Drake, D., Liu, X. S., Bennett, D. A., Colaiácovo, M. P., and Yankner, B. A. (2014). REST and stress resistance in ageing and Alzheimer's disease. *Nature*, 507(7493):448–454.

Manel, S., Schwartz, M. K., Luikart, G., and Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, 18(4):189–197.

Manolio, T. A., Brooks, L. D., and Collins, F. S. (2008). A HapMap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation*, 118(5):1590–1605.

Mardia, K., Kent, J., and Bibby, J. (2009). *Multivariate Analysis*. Academic Press.

McArdle, B. H. and Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297.

McVean, G., Spencer, C. C. A., and Chaix, R. (2005). Perspectives on Human Genetic Variation from the HapMap Project. *PLOS Genetics*, 1(4):e54.

Milan, L. and Whittaker, J. (1995). Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 44(1):31–49.

Morrison, D. F. (2005). Multivariate analysis of variance. *Encyclopedia of biostatistics*, 5.

Neyman and S., J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. (Tlanslated and edited by D. M. Dabrowska and T. P. Speed, Statistical Science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2017). *vegan: Community Ecology Package*. R package version 2.4-5.

Præstgaard, J. T. (1995). Permutation and bootstrap kolmogorov-smirnov tests for the equality of two distributions. *Scandinavian Journal of Statistics*, pages 305–322.

R Core Team (2016). R: A Language and Environment for Statistical Computing.

Rotimi, C., Leppert, M., Matsuda, I., Zeng, C., Zhang, H., Adebamowo, C., Ajayi, I., Aniagwu, T., Dixon, M., Fukushima, Y., Macer, D., Marshall, P., Nkwodimmah, C., Peiffer, A., Royal,

C., Suda, E., Zhao, H., Wang, V. O., and McEwen, J. (2007). Community Engagement and Informed Consent in the International HapMap Project. *Public Health Genomics*, 10(3):186–198.

Roy, S. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, 24(2):220–238.

Seber, G. A. (2009). *Multivariate observations*. John Wiley & Sons.

Skipper, M. (2007). Genomics: HapMap Phase II unveiled. *Nature Reviews Genetics*, 8(11):827–827.

Smyth, D. J., Cooper, J. D., Bailey, R., Field, S., Burren, O., Smink, L. J., Guja, C., Ionescu-Tirgoviste, C., Widmer, B., Dunger, D. B., Savage, D. A., Walker, N. M., Clayton, D. G., and Todd, J. A. (2006). A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (*IFIH1*) region. *Nature Genetics*, 38(6):617–619.

Tang, Z.-Z., Chen, G., and Alekseyenko, A. V. (2016). PERMANOVA-S: association test for microbial community composition that accommodates confounders and multiple distances. *Bioinformatics*, 32(17):2618–2625.

ter Braak, C. J. F. (1992). Permutation Versus Bootstrap Significance Tests in Multiple Regression and Anova. pages 79–85. Springer, Berlin, Heidelberg.

Terwilliger, J. D. and Hiekkalinna, T. (2006). An utter refutation of the 'Fundamental Theorem of the HapMap'. *European Journal of Human Genetics*, 14(4):426–437.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., Stacey, S. N., Bergthorsson, J. T., Thorlacius, S., Gudmundsson, J., Jonsson, T., Jakobsdottir, M., Saemundsdottir, J., Olafsdottir, O., Gudmundsson, L. J., Bjornsdottir, G., Kristjansson, K., Skuladottir, H., Isaksson, H. J., Gudbjartsson, T., Jones, G. T., Mueller, T., Gottsäter, A., Flex, A., Aben, K. K. H., Vegt, F. d., Mulders, P. F. A., Isla, D., Vidal, M. J., Asin, L., Saez, B., Murillo, L., Blondal, T., Kolbeinsson, H., Stefansson, J. G., Hansdottir, I., Runarsdottir, V., Pola, R., Lindblad, B., Rij, A. M. v., Dieplinger, B., Haltmayer, M., Mayordomo, J. I., Kiemeney, L. A., Matthiasson, S. E., Oskarsson, H., Tyrfingsson, T., Gudbjartsson, D. F., Gulcher, J. R., Jonsson, S., Thorsteinsdottir, U., Kong, A., and Stefansson, K. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature*, 452(7187):638–642.

Van Aelst, S. and Willems, G. (2011). Robust and efficient one-way manova tests. *Journal of the American Statistical Association*, 106(494):706–718.

Vicente-Villardón, J. L. (2018). *MultBiplotR: MULTivariate Analysis Using BIPLOTs*. R package version 18.2.09.

Xu, J. and Cui, X. (2008). Robustified MANOVA with applications in detecting differentially expressed genes from oligonucleotide arrays. *Bioinformatics*, 24(8):1056–1062.

Zhang, B. and Srihari, S. N. (2003). Properties of binary vector dissimilarity measures. In *Proc. JCIS Int'l Conf. Computer Vision, Pattern Recognition, and Image Processing*, volume 1.